Study on the Quality of Experience Evaluation Metrics for Astronaut Virtual Training System

Xiangjie Kong(✉), Yuqing Liu, and Ming An

National Key Laboratory of Human Factors Engineering,
China Astronaut Research and Training Center, Beijing, China
kxj2012@outlook.com

Abstract. With the development of virtual reality (VR) technology, it is possible to train astronauts using VR. To make the system more efficient, it is necessary to study the quality of experience (QoE) of astronauts in the virtual environment (VE). Based on the characteristics of virtual training system and the needs of astronauts training, a set of metrics consisting of five higher-level metrics and fifteen lower-level metrics were put forward for the QoE evaluating of the system. In addition, the weight of each higher-level metrics is obtained using analytic hierarchy process (AHP) method. The results of this paper can be used directly in the QoE evaluation of astronaut virtual training system in a quantitative way.

Keywords: Astronaut virtual training system · Quality of experience · Metrics

1 Introduction

Currently, human-centered design is becoming more and more popular. To meet the needs of users is one of the most concerned problems of the developers, and the improvement of the quality of experience has also been paid more and more attention by the developers. The QoE was initially defined as "the overall acceptability of an application or service, as perceived subjectively by an end-user" by the International Telecommunication Union (ITU) in 2007 [1], while the most widely accepted definition was proposed by the Qualinet White Paper in 2012 [2], which is "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state." This definition has been adopted in 2016 by the International Telecommunication Union [3].

After years of development, virtual reality technology has been applied in various fields, providing a new interactive environment and bringing a totally new experience to the users. We believe VR has very good application prospect with the characteristics of flexibility and safety. In order to provide users with better service and enhance user QoE in virtual environment, it is very necessary to study the quality of experience of virtual reality system.

J. Y. C. Chen and G. Fragomeni (Eds.): VAMR 2018, LNCS 10909, pp. 416–426, 2018.
https://doi.org/10.1007/978-3-319-91581-4_31

1.1 Related Work

In 2007, the ITU proposed the concept of quality of experience on the basis of quality of service (QoS). Because QoS is a concept in the field of communication, the research on QoE is mainly concentrated in the network and related fields for a long time, such as multimedia streaming services, IPTV, VoIP, interactive network games and so on. As research progresses, the field of QoE becomes more and more broad. At present, the most widely accepted definition of QoE is the definition given by Qualinet in 2012, and the definition has no longer focused on the field of quality of experience, but the user's subjective feelings while using the product or service. As can be seeing from the change of the QoE definition, QoE is very suitable for describing users' subjective feelings about products or services. Therefore, QoE was chosen to evaluate the astronaut virtual training system from the users' point of view.

A great deal of research has been done on the QoE evaluation of the traditional multimedia applications and services, such as video services and audio services. Especially audio services, methods for QoE evaluation of audio services are fully developed. For example, PSQM (Perceptual Speech Quality Measure) and PESQ (Perceptual Evaluation of Speech Quality) proposed by ITU-T P.861 [4] and P.862 [5]. In terms of visual quality assessment, Peak Signal to Noise Ratio (PSNR) is a traditional evaluation index, but it does not take the visual masking phenomenon into consideration. That is to say, every single pixel error affects the value of PSNR, even though the error may not be noticeable. So, a method called MPQM (Moving Pictures Quality Metric) which incorporates human vision characteristics was proposed for video quality assessment [6]. Visual information fidelity (VIF) [7], structure similarity index (SSIM) [8] and texture similarity [9] are several commonly used metrics for visual quality evaluation, which are resulting from the comparison of input and output signals frame-by-frame. Janssen et al. used two metrics, naturalness and colorfulness, to evaluate the visual quality of images and pointed out that people showed a clear preference for more colorful, yet slightly unnatural images [10]. In the FUN model, three metrics, Fidelity, Usefulness, and Naturalness were used to evaluate the visual quality, and the model was considered as a milestone in the road that took visual quality to be evolved into QoE [11]. Zhang and Kuo proposed a model called GLS from the respect of distortion and information loss for the evaluation of QoE on retargeted images and pointed out GLS quality index has stronger correlation with human QoE than other existing objective metrics in retargeted image quality assessment [12]. Three metrics were analyzed in the model. They are global structural distortion (G), local region distortion (L) and loss of salient information (S). In addition, some QoS indicators such as packet loss, delay, jitter, bandwidth, buffer time, and buffer rate are also commonly used for QoE evaluation in network-related applications.

3D images and videos are regarded as milestone in the multimedia technologies, on the other hand, new challenges are introduced in 3D stereoscopic image and video quality assessment compared with traditional multimedia and just a little research has been down on the 3D QoE evaluation. Based on the symmetry of binocular visual perception, Qi et al. proposed a stereoscopic image QoE assessment model [13]. They trained a SVR model for QoE prediction using the HOG features extracted from the two

views' image separately. Chen et al. proposed a linear model with three metrics, Image quality, depth quantity and visual comfort, for 3DTV QoE evaluation [14]. Two experiments were designed to determine the weights of the parameters and they are working on a more general model. Similar with Chen's work, Kazuhisa et al. using video quality, depth quality, discomfort and fatigue as the metrics for 3D video QoE evaluation [15]. All their work promotes the study of 3D QoE, while it is not enough.

As the study gets deeper and more extensive, study on QoE is not limited to audio-visual aspects anymore. As described in the Qualinet model, QoE is a multidimensional quality that can be decomposed in a set of perceptual attributes called features, that is perceivable, recognized and namable, and these features can be classified into four categories: features at the level of perception, at the level of interaction, at the level of usage, and at the level of service. As a new media technology, VR technology not only possesses stronger visual expression than traditional media, but also provides new and diversified interaction modes and stronger interaction levels. Some research has been done on VR QoE, for example, Keighrey et al. compared the user QoE of an interactive and immersive speech and language assessment implemented in both AR and VR [16]. Both objective metrics of heart rate and electrodermal activity and subjective metrics of nature of interaction, immersion, discomfort and enjoyment were considered in their study. The findings of the study demonstrate similar QoE ratings for both the AR and VR environments but users acclimatized to the AR environment more quickly than the VR environment. In Hamam's research, five metrics were used for the QoE assessment model, they are Media Synch, Fatigue, Haptic Rendering, Degree of Immersion and User Intuitiveness [17]. Research on VR QoE is just in its infancy and no well-established evaluation methods and standards have been established. Despite the fact QoE is important for both developers and users. Under these circumstances, Huawei initiated the QoE for VR (Quality of Experience for Virtual Reality) project proposal at the ITU-T SG in January 2017, which means a significant step forward for the community in VR QoE evaluation.

1.2 Astronaut Virtual Training System

With the continuous progress of China space station plan, some new needs of astronaut training arose. Problems such as disorientation and the lack of navigation skills in the space station have to be solved urgently. Traditional physical simulators are useful, but they cannot meet all the needs of astronaut training, especially the simulation of micro-gravity environment. VR technology, which has the characteristics of multi-perception, presence, interaction and autonomy, breaks through the limitations of the environment and simulates the micro-gravity environment visually in an immersive way, has the advantages of flexibility and safety, is an alternative and is becoming an important way for astronaut training. The Astronaut Center of China (ACC) has developed an Astronaut Virtual Training System (AVTS) based on VR, which simulates the conceptual structure and internal environment of the Space Station. The AVTS aims to facilitate new and novel experiences for trainees above and beyond what are possible with traditional physical simulators.

Currently, the virtual training system consists of several modules. The scene in the virtual module is shown in Fig. 1.

Fig. 1. Virtual environment in the module

The astronaut virtual reality training system consists of human-computer-interaction interface, simulation software, virtual space environment and astronauts and trainers. The system architecture is shown in Fig. 2. The system uses head-mounted display (HMD), 3D mouse and position tracker as the main interaction equipment. The astronaut virtual training system was designed to conduct training tasks like navigation, instrumentation and objects operation, environment familiarization, extravehicular roaming and so on. During the training process, the trainees sit at the training platform and keep upright. They can control the avatar's movement by the 3D mouse. The data-glove is used to capture the movements of their hands so they can control the avatar's hand movements. With the HMD, trainees can observe the internal environment of the virtual space station by rotating the head.

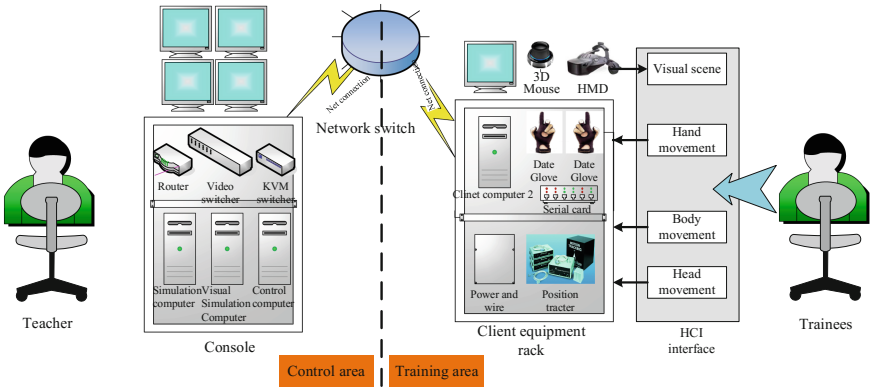

Fig. 2. Astronaut virtual training system architecture

No doubt the VE has its advantages on astronaut training, but the influence of the new technology on astronaut's quality of experience is not considered adequately so far. Based on prior research on QoE and the unique characteristic of the AVTS, a set of metrics is proposed to evaluate the QoE of the AVTS.

2 Method

The easiest and most effective way for QoE measurement is user feedback. However, compared with traditional media quality evaluation, user focuses on more aspects in VE. It is difficult for users to give a reasonable evaluation directly, so the problem needs to be decomposed first. There is no uniform method and metrics for the QoE evaluation until now. Although different organizations established different standards, they all have different focuses. For example, QoE Management Framework (QMF) proposed by ITU-T focused on the evaluation, measurement, basic requirements and preliminary framework of QoE [18], while Customer Experience Management (CEM) framework proposed by TM Forum focused on the definition of CEM, influencing factors and user experience quality model [19]. Obviously, the study on VR QoE is just at its early stage. The existing frameworks or models may not applicable to this new media format and the interactive environment. Coupled with the particularity of astronaut virtual training system, new QoE evaluation metrics need to be built. To the best knowledge of the author, this is the first study on the QoE for a specific VR training system. Based on the definition of QoE and commonly used methods and principles on metrics designing, we proposed five metrics for QoE evaluation, then each metric was further decomposed into multiple lower-level metrics. At the same time, AHP was used to determine the weight of the higher-level metrics.

2.1 Metrics Designing

In order to conduct a reasonable and correct evaluation, we followed four design principles while designing the metrics, which were comprehensiveness, professionalism, rationality and feasibility respectively.

Comprehensiveness means that the study is conducted from different perspectives and the metrics should cover all aspects of the evaluation of QoE. Based on the influence factors of QoE, we evaluate the QoE of AVTS from two aspects of system parameters and user experience. The system parameters refer to the measurable, improved and guaranteed hardware and software characteristics of the system; user experience is the user's subjective experience of the system, including three levels of perception, physiology and psychology.

Professionalism means that the system's specific characteristics must be taken into consideration while metrics designing. In our study, we consider the characteristics not only of VR technology, but also of training system.

Rationality means that the dimension of evaluation is reasonable, and the metrics selection, information collection and coverage range of the metrics have scientific basis.

We put forward the metrics based on existing methods in the literature, and the system characteristics are fully considered.

Feasibility means that the corresponding parameters of the metrics are accessible, and the calculation method is not complicated.

Based on the features of QoE and the application requirements of astronaut virtual training system, a set of metrics for QoE evaluation of astronaut virtual training system are proposed under the consideration of the principle mentioned above. We will evaluate the QoE of AVTS from5 aspects,and they are named functionality, rendering quality, interactivity, side effects and user intuitiveness respectively. Each aspect consists of several lower-level metrics and the lower-level metrics will be obtained through questionnaires. The complete metrics is showed in Table 1, including five higher-level metrics and fifteen lower-level metrics.

Table 1. AVTS QoE evaluation metrics

	The higher-level metrics	The lower-level metrics
QoE	Functionality	Learnability
		Ease of use
		Usefulness
	Rendering quality	Graphics
		Audio
		Cross modality
		Sensory substitution
	Interactivity	Naturalness
		Validity
		Collaboration
	Side effects	Fatigue
		Cybersickness
	User intuitiveness	Presence
		Confusion
		Emotion

2.2 Metrics Interpreting

The interpretation of the metrics is as follows:

Functionality: Functionality is a costumed parameter mainly focus on the acquisition of cognition and skills. It is then subdivided into four parts: Learnability, ease of use, usefulness, and serviceability. Learnability is the measure of the degree to which a user interface can be learned quickly and effectively, that is, the user feels that the interface is familiar and the operation is easy to remember. Ease of use refers to effectiveness at achieving a person's goals in a way that is satisfying to that person, and is fast and error-free. The input and output devices of virtual training system are complex, and task environment of astronaut is quite different from the daily life environment, so there will be more requirements on the users' skills and cognition. It is necessary to design an easy-to-use system so that the astronaut can complete the task better. Usefulness is the

extent to which the system actually helps to solve users real, practical problems. A system may be easy to use but not relevant to the actual needs of a user. In our study, the users are expected obtain the appropriate knowledge and skills through training. Here, Learnability and ease of use are two common metrics for system evaluation, while usefulness is a metric put forward with the system characteristics taken into account, that is, the system is designed to train astronauts to acquire relevant skills.

Rendering quality: The rendering quality relates to the quality of the two major modalities, namely: graphics and audio. The visual modality is emphasized because the visual perception is at the dominant position in VE. Moreover, cross-modality and sensory-substitution were also considered. The rendering quality includes not only the system parameters such as the field of view, the resolution and the delay of the HMD, but also the perception parameters such as the feeling of reality and the depth information of the three-dimensional objects in the virtual environment. In addition to visual, auditory information also plays an important role. Although there is a certain tolerance for the non-synchronization of audiovisual, users will be confused if they exceed the threshold. The metric cross-modality is mainly design for the synchronization of different modality. In addition, hearing also plays an important role in sensory-substitution. Without tactile in the VE, visual and auditory information is combined to make reasonable judgement such as whether the avatar has pushed the button, while one can judge this through haptic stimulus in real life.

Interactivity: Interactivity is a major feature of VR, we subdivided this parameter into three parts: naturalness, validity, and collaboration. Naturalness refers to whether the interactive operation is natural and true. The validity refers to whether the result of the operation conforms to the user's expectation, and the cooperation refers to whether the user and the interaction interface can cooperate well. With the use of new interactive devices, the lack of sense (such as tactile) and the existence of sensory-substitution, the interaction in VE is quite different from the traditional way. The way of interaction and the interface must be properly designed so the astronaut can migrate the skills obtained in VE more efficiently to real scene.

Side Effects: This parameter mainly evaluates the fatigue and discomfort caused by the use of the system. Fatigue includes visual fatigue caused by wearing the head mounted display (HMD), muscle fatigue caused by manipulation of 3D mouse and mental fatigue caused by the use of the system; discomfort refers to the symptoms of dizziness, nausea and other simulation sickness. Side effect is a universal problem of virtual reality system. Stanley and Kennedy summarize previous studies and found that 80%–95% of participants had varying degrees of side effects when they were roaming in a virtual environment wearing HMD [20]. In addition, fatigue, especially visual fatigue caused by HMD cannot be avoided.

User Intuitiveness: Intuition reflects the user's most instinctive behavior and feelings, including presence, confusion and emotion. Presence is a unique characteristic of VR system and is a manifestation of the overall performance of the system, which is related to the physical characteristics, virtual presentation technologies, interactive features and the type of task presented by the system. A good sense of presence means that trainees feel themselves in a system-built virtual environment just like in the real world. According to Schmidt and Young's theory of transfer of training [21], if the

immersive virtual training environment is similar to the mission environment, virtual training will promote learning and enhance positive transfer, which means that the user's performance in the actual task will be better. Therefore, a good sense of presence not only helps trainees get a better experience, but also helps to improve training performance. Confusion and emotion are the characteristics of the degree of consistency between system response and user expectation. If the operation of the system is inconsistent with the expectation of the user, it will cause confusion; if the system response does not match user's operation, the user will be frustrated and depressed.

2.3 Metrics Weighting

The weight of metrics is the characterization of the degree of importance of metrics. Several methods have been used for metrics weighting in prior research, such as Delphi, AHP, principal component analysis(PCA), artificial neural network and so on.

After the metrics is constructed, the AHP is used to determine the weight of the higher-level metrics. The AHP is a theory of measurement through pairwise comparisons and relies on the judgements of experts to derive priority scales [22]. The comparisons are made using a scale of absolute judgements that represents, how much more, one element dominates another with respect to a given attribute. Table 2 exhibits the scale.

Table 2. The fundamental scale of absolute numbers

Intensity of importance on an absolute scale	Definition	Explanation
1	Equal importance	Two activities contribute equally to the object
3	Moderate importance of one over another	Experience and judgement slightly favour one activity over another
5	Essential or strong importance	Experience and judgement strongly favour one activity over another
7	Very strong or demonstrated importance	An activity is favoured very strongly over another; its dominance demonstrated in practice
9	Extreme importance	The evidence favouring one activity over another is of the highest possible order of affirmation
2, 4, 6, 8	Intermediate values between the two adjacent judgments	When compromise is needed
Reciprocals	If activity *i* has one of the above non-zero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i*	
Rationals	Ratios arising from the scale	If consistency were to be forced by obtaining *n* numerical values to span the matrix

Four experts in the fields of virtual reality and human-computer interaction participating in rating the importance of the metrics. Finally, we have the following for the matrix of pairwise comparisons of the metrics with respect to the overall focus, shown in Table 3.

Table 3. Pairwise comparison matrix for higher-level metrics

	Functionality	Rendering quality	Interactivity	Side effects	User intuitiveness
Functionality	1	3	3	2	3
Rendering quality	1/3	1	1	2	2
Interactivity	1/3	1	1	3	2
Side effects	1/2	1/2	1/3	1	1/3
User intuitiveness	1/3	1/2	1/2	3	1

To decide whether the comparison is consistent or not, a consistency test is performed on the judgment matrix. The consistency ratio (CR) helps judge the consistency in pairwise comparison.

For a n-order matrix with the largest eigenvalue of λ_{max}, consistent index (CI) that describes deviation or degree of consistency of the judgment matrix is computed by the following formula:

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{1}$$

Then, the consistency ratio (CR) is defined as the following formula:

$$CR = \frac{CI}{RI} \tag{2}$$

Where *RI* is called mean random consistency index, and the value of *RI* is showed in Table 4. The comparison is considered consistently if $CR < 0.1$.

Table 4. Mean random consistency index

Order	1	2	3	4	5	6	7	8	9
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

The results of the test through calculations showed that $\lambda_{max} = 5.3595, n = 5, CI = 0.0899$ and $CR = 0.080 < 0.1$, so we believe the comparison is consistent.

Finally, the weight of each higher-level metrics through calculations is shown in Table 5.

Table 5. The weight of each higher-level metrics

Functionality	Rendering quality	Interactivity	Side effects	User intuitiveness
0.395	0.180	0.197	0.091	0.136

As shown in the result, functionality has the highest weight of 0.395, which is in line with our expectation. Functionality corresponds to the features of QoE of usage and service. The largest weight also manifests that the function of the astronaut virtual training system is taken seriously. Rendering quality and user intuitiveness corresponding to perception features of QoE, and QoE is characterized from the aspects of system and user respectively. Interactivity is a major aspect of the virtual training system that is different from traditional media. The weight of interactivity is 0.197 and is closed to that of rendering quality, indicating that rendering quality and interactivity are equally important. They two both play important roles in guaranteeing the sense of reality of VE. The weight of side effects and user intuitive were 0.091 and 0.136, respectively, which were relatively low because side effects could be suppressed by reasonably arranging training sessions, and the weight of user intuitiveness should not too large due to individual differences. Therefore, the weight of the index is relatively reasonable. However, due to the fact that the research on the QoE for VR has just begun, there may be some inaccuracies in the expert's experience, and the structure and weight of the metrics will be further optimized in subsequent research, and a set of experiments will be conducted for this work.

In this paper, only the weight of higher-level metrics was calculated. Because the lower-level metrics may be adjusted in practical application, and the processing methods should be different.

3 Conclusion

In this paper, we presented a set of evaluation metrics, consisting of five higher-level metrics and fifteen lower-level metrics, for the QOE evaluation, which makes an evaluation from multiple dimensions. System characteristics and the features of VE are fully considered so that the metrics is applicable for the QoE evaluation of the AVTS. In addition, the weight of each higher-level metrics is obtained through the analytic hierarchy process, which can be used for the QoE evaluation of the system. The result of our study can also be used as a reference for the QoE evaluation of virtual training system in other fields.

References

1. ITU-T: Definition of quality of experience (QoE). International Telecommunication Union, Liaison Statement, Ref: TD 109rev2 (PLEN/12) (2007)
2. Le Callet, P., Möller, S., Perkis, A. (eds.): Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Version 1.2, Lausanne, Switzerland (2013)

3. ITU-T Recommendation P.10: Vocabulary for performance and quality of service, Amendment 5 (2016)
4. ITU-T P.861: Objective quality measurement of telephone-band (300-3400Hz) speech codec (1998)
5. ITU-T P.862: Perceptual evaluation of speech quality (PESQ) (2001)
6. Wang, Y.: Survey of objective video quality measurements. In: International Conference on Intelligent Information Processing, Security and Advanced Communication (2006)
7. Sheikh, H., Bovik, A.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
8. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
9. Zujovic, J., Pappas, T.N., Neuhoff, D.L.: Structural texture similarity metrics for image analysis and retrieval. IEEE Trans. Image Process. Publ. IEEE Sig. Process. Soc. **22**(7), 2545–2558 (2013)
10. Janssen, T.J.W.M., Blommaert, F.J.J.: Image quality semantics. J. Imaging Sci. **41**(5), 555–560(6) (1997)
11. Ridder, H.D., Endrikhovski, S.: image quality is fun: reflections on fidelity, usefulness and naturalness. In: SID Symposium Digest of Technical Papers, vol. 33, no. 1, pp. 986–989 (2002)
12. Zhang, J., Kuo, C.C.J.: An objective quality of experience (QoE) assessment index for retargeted images. In: Proceedings of ACM International Conference on Multimedia. ACM (2014)
13. Qi, F., Zhao, D.: Quality of experience assessment for stereoscopic image based on the symmetry of binocular visual perception. Intell. Comput. Appl. **6**(4), 72–74 (2016)
14. Chen, W., Fournier, J., Barkowsky, M., Callet, P.L.: Quality of experience model for 3DTV. IS&T/SPIE Electron. Imaging **8288**, 2978–2982 (2012). International Society for Optics and Photonics
15. Kazuhisa, Y., Taichi, K., Kimiko, K.: QoE assessment methodologies for 3D video services. NTT Tech. Rev. **11**(5), 1–5 (2013)
16. Keighrey, C., Flynn, R., Murray, S., Murray, N.: A QoE evaluation of immersive augmented and virtual reality speech & language assessment applications. In: Ninth International Conference on Quality of Multimedia Experience. IEEE (2017)
17. Hamam, A., Saddik, A.E.: Evaluating the quality of experience of haptic-based applications through mathematical modeling. In: IEEE International Workshop on Haptic Audio Visual Environments and Games, pp. 56–61. IEEE (2012)
18. ITU-T: ITU-T Recommendation P.10/G.100 "Amendment 2: New Definitions for Inclusion in Recommendation ITU-T P.10/G.100". International Telecommunication Union, Geneva, Switzerland (2008)
19. Customer Experience Management: Driving Loyalty and Profitability TMForum, NJ, USA (2008)
20. Stanney, K.M., Kennedy, R.: Simulation sickness. In: Vincenzi, D.A., Wise, J.A., Moulous, M., Hancock, P.A. (eds.) Human Factors in Simulation and Training, pp. 117–128. CRC Press, Boca Raton (2008)
21. Schmidt, R.A., Young, D.E:. Transfer of movement control in motor skill learning. Transf. Learn. Contemp. Res. Appl. 47–79 (1987)
22. Saaty, T.L.: Decision making with the analytic hierarchy process. Int. J. Serv. Sci. **1**(1), 83–98 (2008)