High-Precision Automatic User Experience Evaluation of 3D Virtual Environments Using Visual-Complexity Image Features

Jarosław Andrzejczak[1,2][0000−0002−4124−0110], Hubert Łabuda[1], and Rafał Szrajber[1,3][0000−0003−2777−0251]

[1] Institute of Information Technology, Lodz University of Technology, al. Politechniki 8, 93-590 Lodz, Poland
http://it.p.lodz.pl
[2] jaroslaw.andrzejczak@p.lodz.pl
[3] rafal.szrajber@p.lodz.pl

Abstract. User Experience testing is an expensive task in terms of time and money due to the need to involve many users. We present a solution to this problem in the form of the automatic evaluation method using image analysis to estimate user impressions of a VR space with high accuracy (significant very strong correlation, Pearson's $r = 0{,}89$ at best). A set of visual complexity image features that describe the image as it might be perceived by a viewer was used: color complexity, object density, directionality, and texture features. Very high correlation values were observed across all level design stages and variants (Pearson's r higher than 0,85 in most cases). As a result, our evaluation method became universal as it can be used even at early stages of virtual space design (from blockout to models with monochromatic materials) and with changes to atmospheric conditions, lighting, and materials.

Keywords: image analysis · Virtual Reality · automatic evaluation · Visual Complexity · Usability testing · User Experience · Impression Curve · level design

1 Introduction

In previous studies, we demonstrated that it is possible to estimate user impressions of a VR space with high accuracy (significant very strong positive correlation, Pearson's $r = 0{,}82$ at best) using automatic evaluation based on image analysis. However, early stages of virtual space design (from blockout to models with monochromatic materials) and in cases involving changes to atmospheric conditions, lighting, and materials exhibited weaker correlation results compared to the final versions of the space. Thus, the automatic evaluation method was sensitive to changes in geometry and lighting. The motivation for this research was to address this issue and explore ways to make the method more universal.

We sought measures that describe the image as it might be perceived by a viewer. For example, whether it is a diverse, colorful scene with a lot of distractors (a street in the center of a large city during rush hour) or rather a static,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

monotonous scene (a desert stretching to the horizon). We decided on the use of a set of eight image features from the Visual Complexity group: color complexity, object density, directionality, and texture features. The automatically computed values of these features will be compared with user ratings of the virtual space obtained using the Impression Curve[1], as it aggregates convergent impressions from different users [2]. Thus, the purpose of the research was to verify the existence of the correlation between the eight Visual Complexity image features (gathered using automatic image analysis) and the Impression Curve (obtained during previous studies conducted on 112 people).

Ultimately, the developed method is intended to serve as the foundation for a tool to automatically estimate user impressions in virtual space, particularly pacing (the general order and rhythm of activities and events in a level) in video games. Such a tool would assist designers in analyzing this from the early stages and, additionally, reduce the cost of UX testing, as there would be no need to test with users each time.

The contributions to research concerning automatic evaluation of the immersive Virtual Reality space, especially in case of the Impression Curve estimation presented in this article, are:

- Improvement in the accuracy of Virtual Reality space UX automatic evaluation through the use of a set of eight Visual Complexity image features.
- Solution for the issue of weaker correlation results between image features and the Impression Curve for the VR space at early stages of virtual space design (from blockout to models with monochromatic materials) and in cases involving changes to atmospheric conditions, lighting, and materials.
- Analysis of the usability of each of the eight Visual Complexity image features individually and in combination into sets, depending on the level design stages and changing factors of the 3D space.

We start with a related work overview in the domain of Visual Complexity features extraction using image analysis in the next section. Then we describe hypotheses and an evaluation method. Next, both test results and their discussion will be presented, as well as observations about data gathered. Finally, ideas for further development and final conclusions will be given.

2 Visual complexity

The motivation behind this research was to expand the image features analyzed in previous works [1] by incorporating visual complexity in the context of human image perception. Earlier studies primarily focused on features related to descriptive statistics (entropy, skewness, and kurtosis) [8], while color contrast

[1] Impression Curve is a measure of the visual diversity and attractiveness of a game level. It assesses the subjective attraction of a given space. For detailed information about the Impression Curve, its acquisition method, and its strengths and weaknesses in the domain of 3D space evaluation, please refer to [2].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

and luminance-based features [12] were used in the form of average measures for the entire image frame. At the same time, features based on saliency and motion maps (such as balance and density) [10] demonstrated some of the highest correlations across all analyzed cases, prompting us to pursue research in this direction.

As we sought measures that describe the image as it might be perceived by a viewer, we selected visual complexity features in several contexts: in the domain of color (using advanced features related to color rather than only analyzing brightness and contrast), as well as histogram distribution, image texture and patterns, variability/monotonicity of the image (including the number of objects within the image), and its potential directionality.

In this study, **three color metrics** were applied. All of them operate in the CIELab color space. The first two were chosen from [6], where the authors conducted an experiment to identify combinations of different color metrics that correlated most strongly with user ratings. Two of the highest-performing metrics were selected for this study. The first metric combines the maximum and minimum standard deviations in the ab color space with the mean Chroma value for a given color. The second metric is described as the length of a vector composed of the standard deviations in the ab color space and the mean Chroma value for the given color. The third metric applied in this study was presented as one of many used to predict users' first impressions of website aesthetics [14]. It is described as the sum of the mean saturation value and its standard deviation, where saturation is calculated as Chroma divided by lightness in the CIELab color space.

Based on [13], two additional image parameters were selected that can be used to predict user impression ratings: the **number of colors after posterization** and the **standard deviation of grayscale** values. The authors of mentioned work indicate posterization into six levels as the most appropriate approach—this was also applied in this study. The process of finding unique colors after posterization involves iterating over each pixel in the RGB image, posterizing the value according to the predefined number of classes, and adding this value to the corresponding class object. The standard deviation of grayscale values was calculated based on an image converted to grayscale.

A feature of an image that can be used to describe the quantity of objects in the image is the **number of edges**. In [11], this metric was used to determine the amount of information contained in the image and its readability. Meanwhile, [5] proposed using the number of edges as one of the parameters defining the visual complexity of images. First, the image frame is converted to grayscale and blurred using a Gaussian filter with a kernel size of 5×5. Next, edge detection is performed using the Canny algorithm. Finally, the value of this parameter is calculated as the ratio of white pixels to the total number of pixels in the image.

According to [4], one of the important features of an image for human perception is its **directionality**. Based on this parameter, it is possible to determine whether the elements in the image are aligned in the same direction or whether the image is chaotic (Fig. 1). To compute the directionality value, an algorithm

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

described in [17] was applied. It involves calculating two directional derivatives for each pixel in the image and then verifying whether the mean value of both derivatives exceeds a threshold value. The obtained derivative values are subsequently converted into angles. These angles are grouped into sixteen bins. Peaks are identified within these bins, and the final result is calculated as the sum of the squared differences between the value of each bin belonging to a peak and the peak value.

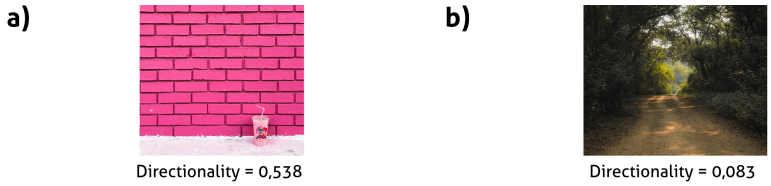

Fig. 1. Images illustrating the consistency of the directionality measure with human perception. (a) Image with high directionality, with a value of 0,538. (b) Image with low directionality, with a value of 0,083.

The final feature chosen for analysis to assess the visual complexity of an image is **coarseness** [17]. It refers to a textural feature that describes the perceived roughness or granularity of a surface or pattern in visual perception. The algorithm involves calculating the mean pixel value within increasingly larger squares, in both vertical and horizontal directions, around the currently considered pixel. The largest possible square size in either direction, for which the difference in mean values was the greatest, is then selected. The final coarseness value is computed as the mean of these selected square sizes.

In summary, the following eight image features were used in this study for automatic evaluation: three color metrics in the CIELab color space, standard deviation of shades of gray, number of colors after posterization, edge ratio, directionality, coarseness.

3 Optimal parameters of the methods used

Two of the features used in this research — edge ratio and coarseness — required preliminary experiments to determine the optimal parameters (yielding the highest correlation values with the Impression Curve) for their calculation algorithms. These experiments are described in the following subsections.

3.1 Edge ratio feature parameters selection

The values of the edge ratio parameters were selected based on an experiment involved calculating the edge count parameter using t_1 and t_2 for all recordings considered in the study. The parameters were computed as follows:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

1. In the outer loop, t_1 took values ranging from 20 to 240, with a step size of 20.
2. In the inner loop, t_2 took values ranging from $t_1 + 10$ to 250, with a step size of 20.
3. For each pair of t_1 and t_2, the edge count parameter value was calculated and recorded in an array.
4. After calculating all values for a given recording, Pearson and Spearman correlations were computed between the obtained results and the impression curve for that recording.

The results showed that the highest possible correlations, as well as the largest number of p-values below 0,05 (Fig. 2), were achieved for $t_1 = 200$ and $t_2 = 250$. High values of t_1 and t_2 indicate that only the most distinct edges are detected. In contrast, for lower values, such as $t_1 = 50$ and $t_2 = 100$, significantly more edges are detected, including many that are not as prominent in the original image (Fig. 3).

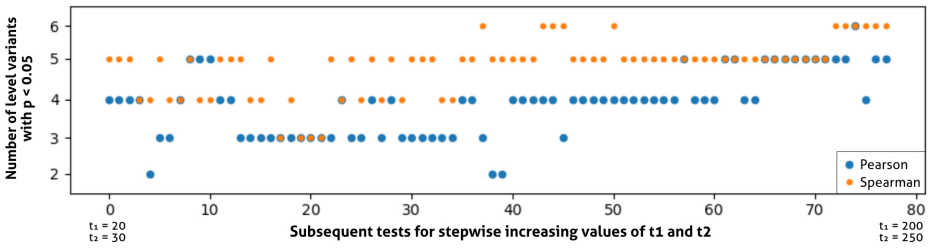

Fig. 2. The numbers of recordings for which the p values for Pearson's correlation (blue) and Spearman's correlation (orange) are less than 0,05. The horizontal axis represents successive iterations calculating the t_1 and t_2 parameters, where 0 means t_1 =20 and t_2 =30, and 78 means t_1 =240 and t_2 =250.

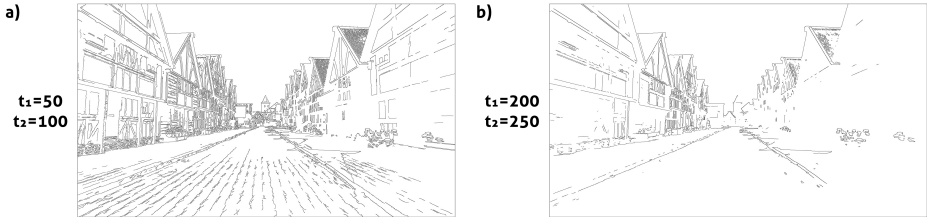

Fig. 3. An example frame of video recording of variant F (final level) with parameters: a) t_1 =50 and t_2 =100 and b) t_1 =200 and t_2 =250.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

3.2 Frame size selection for coarseness feature

The algorithm for calculating the coarseness feature is highly time-consuming due to the large number of pixel array accesses. For this reason, an experiment was conducted to evaluate how the correlation of the parameter value changes with variations in frame size and interpolation algorithm.

The experiment involved analyzing the algorithm's results for 3 seconds of recordings from variants D (level with monochromatic materials), E (level with final materials), F (final level), and G (advanced blockout with modified geometry), for frames of dimensions 1920×1080, 960×540, 480×270, and 240×135. Pearson and Spearman correlations were computed between the obtained results and the impression curve for the same recordings.

The results indicated that the best correlation and the lowest p-value were achieved for frames with dimensions 240×135 when using linear interpolation (Tab. 1).

Table 1. Pearson and Spearman correlation values between the algorithm calculating granularity and the impression curve for a frame size of 240×135 for four selected level design variants.

Level Design Variant	Pearson r	p value	Spearman r	p value
D - models with monochromatic materials	0,55	<0,01	0,49	<0,01
E - models with final materials	0,55	<0,01	0,57	<0,01
F - final level version	0,63	<0,01	0,65	<0,01
G - geometrical changes	0,37	<0,01	0,11	0,32

4 Evaluation

The goal of the evaluation was to verify whether image features from the Visual Complexity domain could improve the correlation of automatic interface evaluation (gathered using automatic image analysis) with the assessment of impressions of the virtual space obtained from users using the Impression Curve. This improvement would enhance the quality of automatic estimation of user impressions of virtual environments. For this purpose, the Pearson and Spearman correlation coefficients were used [9]. All the level design stages, as well as influential factors affecting 3D space impressions (such as lighting condition changes, and geometrical and material changes), described in [2] and [1], were included in the analysis (Fig. 4).

The study was divided into three parts. First, the correlation of the eight individual Visual Complexity image features with the Impression Curve was analyzed. Subsequently, the image features with the highest correlation values were combined into sets and tested again for correlation with the Impression Curve to identify potential improvements in the strength of the correlation. Finally, the results obtained for Visual Complexity features were combined with the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

best-performing features from previous studies [1] to verify whether the inclusion of Visual Complexity features increased the correlation with the Impression Curve, thereby improving the quality of user experience estimation for 3D virtual spaces. Simultaneously, at this stage, the study investigated whether applying Visual Complexity image features improved correlation results for the most challenging design stages and virtual space variants (e.g., initial blockout stages and variants involving atmospheric or lighting changes).

The hypotheses for each part of the study were formulated as follows:

1. **First part:** There exists a significant correlation between individual image features and the Impression Curve for the corresponding VR space.
2. **Second part:** The correlation with the Impression Curve is higher for sets of combined image features compared to individual image features.
3. **Third part:** The correlation with the Impression Curve is higher for combined image feature sets that include Visual Complexity features than for the best-performing image feature sets identified in previous research.

Similar to previous studies on the automatic evaluation of the Impression Curve, controlled 3D space designs were utilized to minimize the influence of external factors such as player actions, gameplay rules and constraints, as well as narrative and lore elements typically present in commercial game designs [1]. The twelve level variants showing successive design stages were used, described in details in [2] (Fig. 4). Particular attention in this study is given to variants A-D (early design stages without textures and final models), lighting condition changes (e.g., day-night transitions: variant L), and atmospheric conditions (e.g., rain, fog: variant W), which in previous studies demonstrated lower correlation results compared to other variants.

Fig. 4. The twelve level variants showing successive design stages used in this study for image analysis and correlation with Impression Curve. A - simple blockout; B - advanced blockout; C - models without materials; D - models with monochromatic materials; E - models with final materials; F - final level version; G - geometrical changes; L - lightening condition changes; W - weather changes; M - material changes; X - expression added; O - extra models added.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

5 Results and analysis

The experiment stages were as follows: first, for each of the video game level variants the Impression Curve data (gathered with users) was interpolated between the measure points to match the frequency of data calculated using image analysis for this level variant walkthrough video; next, the image features were calculated and refined using respectively mean and median for 20 subsequent frame intervals[2]; finally, the Pearson and Spearman correlation between those data were calculated.

The recordings of twelve variants of the video game level variants (used in [2]), including twenty-nine thousand three hundred and thirty-nine frames in total, were analyzed. As a result, eighteen data sets were obtained and used to generate two hundred and twenty-eight correlation plots.

5.1 Individual image features correlation

The first part of the study focused on testing each of the eight Visual Complexity image features individually for their correlation with the interpolated Impression Curve for each of the twelve variants of the video game level described earlier. The results for each feature-variant pair were recorded numerically (Table 2) and visualized as correlation plots for easier analysis (Fig. 5). Each data point in the graph represents the median over a 20-frame interval of the video data. Feature values are plotted in red, while Impression Curve values are plotted in green. The charts also display the Pearson correlation coefficient calculated for the entire Impression Curve and the corresponding p value.

We observed numerous significant correlation values (both positive and negative with a tendency for negative correlations to prevail) between image features and Impression Curve values. Observations ranged from a few weak correlations (r values between 0,20 and 0,29) to moderate correlations in most cases (r values between 0,30 and 0,39), and even over a dozen strong correlations (r values between 0,40 and 0,69). Notably, there was no variant without at least one significantly related image feature, and in most cases, several weak or moderate correlations were identified. Furthermore, certain image features demonstrated a tendency to correlate more frequently, while others exhibited at least a weak correlation only once or twice.

It can be observed that the simple blockout variant (A) achieved correlation values of $p > 0{,}15$ (positive or negative) for all parameters except Directionality. The color metrics showed the highest correlation for the initial level variants in both cases. Additionally, Color Metric 3 achieved strong correlation values not only for the first four stages but also for the variants with local and global changes (L and W). The Edge Ratio image feature correlated best in stages without additional global or local modifications, specifically from the simple

[2] A range of twenty frames was used in this study, similarly to previous research [1], where it was demonstrated that this range shows the highest correlation values with significance $p < 0{,}05$ (in many cases $p < 0{,}01$).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

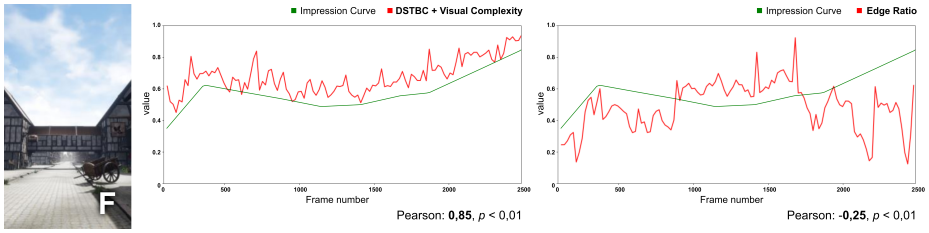

Fig. 5. Correlation plot examples for final level design variant (F variant, on the left). Two image feature correlation plots are presented: one with significant very strong positive correlation - combined score of Density Complexity, Size Complexity, Total Complexity, Balance Complexity (DSTBC for short) and all the Visual Complexity features (Pearson $r = 0{,}85$ with $p < 0{,}01$, center) other with weak linear correlation and - single feature Edge Ratio (Pearson $r = -0{,}25$ with $p < 0{,}01$, right). The combined score presents significantly higher (more than three times higher) correlation value than the single feature separately with significant very strong positive correlation.

blockout (A) to the final level (F). For other variants, where the number of objects in the scene differed from their basic counterparts or edges were less visible — such as in the final level with altered lighting (L) — the correlation decreased. For the advanced blockout stage with modified geometry (G) and the final level with added expression (X), the number of correlating parameters was least of all, with low correlation values.

The results show (Tab. 2) that the introduced image features perform best for the first two stages of virtual space design (variants A - simple blockout and B - advanced blockout) and for atmospheric changes (variant W). It is worth noting that most of these correlations are negative (which indicates that the two variables tend to move in opposite directions). Furthermore, both color-based features (Color Metrics 1 to 3) and the other tested Visual Complexity features, in various combinations, showed correlation values indicating a moderate to strong correlation in at least two out of the first four variants (A to D): Color Metric 2, Posterization, Shades of Gray SD, Edge Ratio, Directionality, and Coarseness. Color Metric 1 demonstrated significant correlation in three subsequent variants, while Color Metric 3 exhibited strong correlations in as many as four consecutive variants.

Those results represent a substantial improvement compared to previous studies and addresses the observed issue of low performance in the automatic estimation of user impressions for the early stages of virtual space design [1].

Since, for most cases, we observe no significant difference between the Pearson and Spearman correlation coefficient values, which suggests a linear nature of the correlation. Thus, in further combined image features, we focused on the Pearson correlation coefficient, as a linear correlation is more desirable for the future automatic evaluation system for video game level design.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

Table 2. Pearson's correlation values for individual image features for all twelve level design variants. Feature values were calculated as median for the intervals of twenty frames. The significant correlation results (with $p =< 0{,}01$) are marked with a grayscale background color (the darker the color, the higher the correlation value). Strong correlation (r value between 0,40 and 0,69) was outlined with a white text color. We can observe a strong correlation values for the early stages of design (first four stages, A to D) and also for the variants with local and global changes (L and W). Combination of features gave even better results (strong and very strong correlation). Best features combined - multiple best feature regression correlations (r value higher than 0,2 and $p =< 0{,}01$) for individual variants. All features combined - multiple regression correlations of all the above visual complexity parameters. A to O - variants described in details in Fig. 4.

	Level Design Variant											
Image Feature	A	B	C	D	E	F	G	L	W	M	X	O
Color Metric 1	-0,26	-0,25	0,40	-0,10	0,16	-0,09	-0,05	-0,12	-0,17	0,16	0,08	0,07
Color Metric 2	-0,39	-0,29	0,05	0,03	0,03	-0,14	-0,12	-0,14	-0,21	0,11	0,13	0,06
Color Metric 3	-0,26	-0,31	0,38	0,38	0,03	-0,02	-0,04	0,23	-0,24	0,11	0,02	0,14
Posterization	-0,40	-0,12	0,09	0,23	-0,07	0,33	-0,14	-0,06	0,33	0,54	0,31	0,61
Shades of Gray SD	-0,51	-0,28	-0,18	0,09	-0,17	-0,23	-0,29	-0,33	-0,26	0,00	0,14	0,03
Edge Ratio	0,39	0,08	-0,20	-0,17	-0,32	-0,25	0,04	0,10	-0,07	-0,07	0,21	0,03
Directionality	0,05	0,01	-0,40	-0,32	-0,20	-0,43	-0,11	-0,48	-0,13	-0,34	-0,16	-0,34
Coarseness	-0,30	-0,31	-0,09	0,03	0,31	0,19	-0,05	-0,36	-0,26	0,24	0,07	0,25
Best features combined	0,74	0,53	0,64	0,49	0,35	0,60	0,29	0,64	0,66	0,60	0,34	0,63
All features combined	0,76	0,58	0,77	0,63	0,59	0,73	0,61	0,73	0,73	0,80	0,55	0,73

5.2 Combined image features correlation

As a next step, the calculated image features were combined into sets using multiple regression to verify whether combining them will increase the correlation with the Impression Curve against individual parameters. This test was also intended to select a set of parameters for later combination with the results of previous studies. *Best features combined* (Table 2) presents the multiple regression correlations of the best parameters (r value higher than 0,2 and $p =< 0{,}01$) for each level design variant. We also checked the multiple regression correlations of all analyzed visual complexity parameters together (*All features combined* in Table 2), which yielded even better correlation values with the Impression Curve for all variants of the tested virtual space.

There was significant strong or very strong positive correlation in all cases. In all cases, the combined feature sets correlated significantly better or at least the same as the single ones included in them (Fig. 5). An interesting observation is that although some features individually did not exhibit significant correlation (either due to a low Pearson's r value or p values $> 0{,}05$), when combined with other features, they positively influenced the overall correlation results (Tab. 2 - the last two rows). This effect is particularly noticeable in variants C (models without materials) and F (final level version), where the increase in Pearson's r value reached 0,13, elevating the correlation from strong to very strong. This

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

leads us to the conclusion that while a given feature may not independently carry significant informational value in the context of UX evaluation for virtual spaces, its combination with other features can provide meaningful insights. This is analogous to health diagnostics—individual symptoms may not clearly indicate a condition, but the presence of multiple symptoms together can provide a strong indication.

Still, local and global geometrical changes (variants G and X) exhibited the lowest correlation results (Pearson's r median: 0,29 and 0,34, respectively) for the set of best-correlating features. However, incorporating all Visual Complexity features significantly increased these values, doubling the correlation in the first case (increase of r from 0,29 to 0,61). Since combining all features also improved correlation in all other cases, it is recommended to use the complete set of Visual Complexity features to achieve the best results, provided that computational cost and time are not limiting factors.

5.3 Gain relative to previous studies

Table 3. Pearson's correlation values for combined image features for all twelve level design variants. Features were combined using multiple regression. The four best correlated image features from previous research: Density Complexity + Size Complexity + Total Complexity + Balance Complexity plus all the Visual Complexity features (called DTSBVC for short) were bases for all four combined sets. The correlation results are color-coded as a heatmap with a grayscale background color (the darker the color, the higher the correlation value). Very strong correlation (r value higher than 0,80) was outlined with a white text color and bold text. All the results were significant ($p =< 0{,}001$). There was significant correlation in all cases, where the most universal combination of image features presents the second and third row. We can observe very strong correlation not only for final design (F) but even for simple blockout (A) and variants with lightning and weather changes (L and W). A to O - variants described in details in Fig. 4.

	Level Design Variant											
	A	B	C	D	E	F	G	L	W	M	X	O
DSTBVC + Average Contrast	**0,83**	0,65	**0,81**	0,79	0,75	**0,87**	0,64	0,79	**0,87**	**0,86**	0,70	0,78
DSTBVC + Average Contrast + Average Saturation + Hue Entropy	**0,87**	0,71	**0,85**	**0,85**	**0,83**	**0,88**	0,69	0,79	**0,87**	**0,89**	0,73	**0,85**
DSTBVC + Average Contrast + Average Saturation + Hue Entropy + Luminosity Kurtosis	**0,87**	0,74	**0,85**	**0,85**	**0,89**	**0,89**	0,69	0,80	**0,87**	**0,89**	0,76	**0,85**
DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy	**0,83**	0,67	**0,83**	**0,81**	**0,83**	**0,86**	0,67	**0,88**	**0,87**	**0,87**	0,70	0,78
DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy + Luminosity Entropy	**0,84**	0,69	**0,83**	**0,84**	**0,83**	**0,88**	0,67	**0,88**	**0,88**	**0,87**	0,71	0,79

As a final step, we combined results of Visual Complexity features with five of the best correlated image feature sets from previous research [1] (Tab. 3). We observe an increase in correlation across all analyzed cases (Tab. 4). Regardless

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

of the tested level variant or the combination of image features used, incorporating Visual Complexity features consistently improved the correlation between automatically obtained values and actual user impression ratings (with Pearson's r increase compared to previous studies by 0,07 to 0,31 in the best cases, with an average improvement of 0,18). Notably, the highest gain was observed for the early stages of level design (variants A to D: simple blockout, advanced blockout, models without materials, models with monochromatic materials) and for variant O (which featured significant geometric variations), as well as for the atmospheric condition changes variant (W) (Fig. 4). In these cases, the median Pearson's r gain exceeded 0,2 in almost every instance, with a small standard deviation between 0,01 and 0,05. The actual Pearson's r correlation values for these variants ranged from 0,65 (strong positive correlation) to 0,87 (very strong positive correlation), with a mean value of 0,80 and a standard deviation of 0,06. Thus, the inclusion of Visual Complexity features resolved the issue of weaker performance for the early stages of virtual space design, which, due to their synthetic and monotonic form or lack of textures (particularly in the first two blockout variants), are less representative of real-world environments than the final variant, making them more challenging for automatic evaluation.

Table 4. Gain in the Pearson's correlation values for combined image features for all twelve level design variants compared to previous studies. Features were combined using multiple regression. The four best correlated image features from previous research: Density Complexity + Size Complexity + Total Complexity + Balance Complexity plus all the Visual Complexity features (called DTSBVC for short) were bases for all four combined sets. The correlation results are color-coded as a heatmap with a grayscale background color (the darker the color, the higher the correlation value). Highest gain (difference higher than 0,20) was outlined with a black background. The largest gain is observed for variants that in earlier studies gave lower correlation scores, that is: the early stages of virtual space design (variants A through D), the variant with additional 3D models (O) and atmospheric changes (W). A to O - variants described in details in Fig. 4.

	Level Design Variant											
	A	B	C	D	E	F	G	L	W	M	X	O
DSTBVC + Average Contrast	0,16	**0,23**	**0,20**	**0,25**	0,12	0,14	**0,24**	0,15	**0,27**	**0,20**	0,19	**0,23**
DSTBVC + Average Contrast + Average Saturation + Hue Entropy	0,19	**0,28**	**0,22**	**0,25**	0,11	0,07	0,16	0,12	**0,25**	0,14	0,08	**0,22**
DSTBVC + Average Contrast + Average Saturation + Hue Entropy + Luminosity Kurtosis	**0,20**	**0,31**	**0,20**	**0,24**	0,06	0,07	0,16	0,09	**0,25**	0,14	0,05	**0,22**
DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy	**0,21**	**0,21**	**0,21**	**0,22**	0,12	0,12	**0,21**	**0,33**	0,10	0,19	0,18	**0,21**
DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy + Luminosity Entropy	**0,21**	**0,21**	**0,20**	**0,22**	0,12	0,10	0,17	0,12	0,07	0,18	0,19	0,13
Median gain	**0,20**	**0,23**	**0,20**	**0,24**	0,12	0,10	0,17	0,12	**0,25**	0,18	0,18	**0,22**
Standard Deviation of gain	0,02	0,05	0,01	0,01	0,03	0,03	0,03	0,10	0,09	0,03	0,07	0,04

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

Moreover, not only the high results for variants involving geometric and atmospheric changes were observed, but the inclusion of Visual Complexity features also reduced the dependence of automatic user impression estimation on micro-scale changes within the virtual space. In other words, the method maintains a high level of estimation accuracy even in dynamic environments, as long as the changes occur on a micro-scale (e.g., rain, wind, or street-level traffic) without altering the overall structure of the space (e.g., destruction of entire buildings or removal of all trees on a street).

It is also worth noting that, while the gain for the final and pre-final level designs (variants E and F) may appear less significant, we still observe an average increase in Pearson's r of 0,1, bringing the correlation values for these variants closer to 0,9 (where a correlation of 1 indicates a perfect positive linear correlation). Analyzing the median gain for each variant (Tab. 4 - last two rows), the median was greater than 0,1 in all cases and exceeded 0,2 in half of the cases.

Regardless of the combination of parameter connections used, the weakest results were still observed for variant G (significant geometric changes) and variant B (advanced blockout). However, it is important to emphasize that at the current stage of research (after incorporating Visual Complexity features), these correlations are now strong or even very strong.

5.4 Best features for different level design stages

The inclusion of Visual Complexity features in all cases resulted in strong or very strong correlations, while the results for different parameter combinations remained consistent across individual variants (Tab. 3). Nonetheless, certain feature sets stand out.

The most universal combination of image features (i.e., the one yielding the greatest improvement across the largest number of variants) was observed for the combination of Visual Complexity features with two specific sets (Tab. 3): DSTBVC[3] + Average Contrast + Average Saturation + Hue Entropy; and DSTBVC + Average Contrast + Average Saturation + Hue Entropy + Luminosity Kurtosis. Since the results for these two sets are similar, we recommend using the first set due to the lower number of required image features, which reduces computational complexity and processing time (as the set excludes the Luminosity Kurtosis feature).

It is worth noting that for the variant involving changes in lighting (L), the best results were achieved with feature combinations containing descriptive statistics features: DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy; and DSTBVC + Hue Kurtosis + Hue Skewness + Saturation Entropy + Luminosity Entropy. The difference in Pearson's correlation r for variant L is 0,08, while achieving the best results for variant W (atmospheric changes). Therefore, these sets could be used to verify these specific cases, i.e., lighting

[3] In all cases, the base set consists of a combination of Density Complexity, Size Complexity, Total Complexity, Balance Complexity, and all the Visual Complexity features (referred to as DTSBVC for short).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

and weather changes. Similarly, in this case, the sets differ by only one feature (Luminosity Entropy), so for resource and computation efficiency, we recommend the smaller set (the first one).

6 Conclusion

In this study, we investigated the feasibility of improving automatic evaluation of user impressions in Virtual Reality (VR) spaces through the incorporation of eight Visual Complexity image features. Our findings confirm that these features significantly enhance correlation results between automatically computed metrics and user-rated Impression Curve values, addressing previous limitations observed in early-stage virtual space designs and scenarios with dynamic environmental changes. Thus, the inclusion of Visual Complexity features reduces the sensitivity of the automatic evaluation method to minor scene alterations while preserving its robustness in capturing user impressions.

Future research will focus on refining feature selection, exploring deep learning approaches for feature extraction such as [3] and [16], as well as validating the method across a broader range of virtual spaces, including dynamic and interactive environments. We are also currently testing various saliency and motion maps [15] that could be used to extract further features. Additionally, the potential for real-time analysis and integration into game development workflows (such as presented in [7]) will be further explored.

To sum up, the proposed method represents a significant step toward the automation of UX evaluation in VR environments, offering a scalable approach that enhances the efficiency of virtual space design while maintaining high correlation with subjective user experience assessments. From a practical perspective, the developed method contributes to the field of game design and Virtual Reality UX evaluation by enabling designers to assess pacing and user engagement from the early stages of development. This approach can also reduce the cost and time required for traditional user testing by providing an automated alternative for iterative evaluation.

References

1. Andrzejczak, J., Jaros, O., Szrajber, R., Wojciechowski, A.: Image features correlation with the impression curve for automatic evaluation of the computer game level design. In: Computational Science – ICCS 2022. Lecture Notes in Computer Science, vol. 13352, pp. 591–604. Springer (2022)
2. Andrzejczak, J., Osowicz, M., Szrajber, R.: Impression curve as a new tool in the study of visual diversity of computer game levels for individual phases of the design process. In: Computational Science – ICCS 2020. Lecture Notes in Computer Science, vol. 12141, pp. 524–537. Springer (2020)
3. Feng, T., Zhai, Y., Yang, J., Liang, J., Fan, D.P., Zhang, J., Shao, L., Tao, D.: Ic9600: A benchmark dataset for automatic image complexity assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8577–8593 (2023)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17

4. Guo, X., Rao, A., Dai, Q., Xu, S.: Visual complexity perception and texture image characteristics. In: 2011 International Conference on Biometrics and Kansei Engineering. pp. 260–265. IEEE (2011)
5. Guo, X., Rao, A., Dai, Q., Xu, S.: Visual complexity assessment of painting images. In: 2013 IEEE International Conference on Image Processing. pp. 388–392. IEEE (2013)
6. Hasler, D., Suesstrunk, S.E.: Measuring colourfulness in natural images. In: Proceedings of SPIE – The International Society for Optical Engineering. vol. 5007, pp. 87–95. SPIE (2003)
7. Kozłowski, K., Korytkowski, M., Szajerman, D.: Visual analysis of computer game output video stream for gameplay metrics. In: Computational Science – ICCS 2020. Lecture Notes in Computer Science, vol. 12141, pp. 538–552. Springer (2020)
8. Kumari, S., Vijay, R.: Image quality estimation by entropy and redundancy calculation for various wavelet families. International Journal of Computer Information Systems and Industrial Management Applications **4**, 27–34 (2012)
9. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. Wiley, 2nd edn. (2017)
10. Magel, K., Alemerien, K.: GUIEvaluator: A metric-tool for evaluating the complexity of graphical user interfaces. In: Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE). pp. 1–6. Knowledge Systems Institute Graduate School (2014)
11. Miniukovich, A., Angeli, A.D.: Quantification of interface visual complexity. In: Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces (AVI '14). pp. 153–160. Association for Computing Machinery, Como, Italy (2014)
12. Peli, E.: Contrast in complex images. Journal of the Optical Society of America A **7**(10), 2032–2040 (1990)
13. Purchase, H.C., Freeman, E., Hamer, J.: An exploration of visual complexity. In: Diagrammatic Representation and Inference, Lecture Notes in Computer Science, vol. 7352, pp. 200–213. Springer, Berlin, Heidelberg (2012)
14. Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Gajos, K.Z.: Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). pp. 2049–2058. Association for Computing Machinery, Paris, France (2013)
15. Rogalska, A., Napieralski, P.: The visual attention saliency map for movie retrospection. Open Physics **16**(1), 188–192 (2018)
16. Saraee, E., Jalal, M., Betke, M.: Visual complexity analysis using deep intermediate-layer features. Computer Vision and Image Understanding **195**, 102949 (2020)
17. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics **8**(6), 460–473 (1978)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_17