

Є. К. Завальнюк; О. Н. Романюк, д-р техн. наук, проф.;
Ю. О. Ковальова; О. Л. Бобко

ЕКСПЕРТНЕ ОЦІНЮВАННЯ ЯКОСТІ ГРАФІЧНИХ ЗОБРАЖЕНЬ

У статті проаналізовано особливості застосування експертних метрик і методів для оцінювання якості зображень. Розглянуто основні аналітичні метрики, що базуються на обробці зображень, і недоліки їх використання для оцінювання якості візуалізації тривимірних сцен. Обґрунтовано необхідність застосування експертних метрик і методів для повноцінного тестування алгоритмів і моделей тривимірного рендерингу. Наведено числові та якісні вимоги до процесу підбору експертів. Проаналізовано методологію збору вхідних даних і особливості розрахунку експертних метрик якості зображень Mean Opinion Score, Difference Mean Opinion Score, Single-Stimulus Continuous Quality Evaluation, Double Stimulus Impairment Scale, Double-Stimulus Continuous Quality Score. Охарактеризовано переваги й недоліки використання експертних методів анкетування, до яких належать метод ранжування, метод нормування, метод попарного порівняння, методи на основі використання шкал Лайкерта, семантичного диференціалу. Продемонстровано використання методу ранжування для визначення ваг характеристик якості зображення, а також для упорядкування тривимірних зображень за рівнем якості рендерингу. Додатково, розглянуто обчислення коефіцієнта конкордації Кендалла для встановлення рівня узгодженості між експертами. Наведено приклад використання шкали Лайкерта для експертного оцінювання окремих аспектів якості рендерингу. Досліджено застосування дисперсійного аналізу для порівняння варіативності наборів оцінок якості графічних сцен. Розглянуто можливість здійснення групової експертизи, зокрема, на основі методу Дельфі, для оцінювання результатів застосування особливо складних алгоритмів тривимірної візуалізації. Наведені у статті методи й метрики призначені для оцінювання якості формування зображень у високопродуктивних і фотореалістичних системах рендерингу.

Ключові слова: *еталонне зображення, тестове зображення, Mean Opinion Score, Double Stimulus Impairment Scale, коефіцієнт конкордації Кендалла, метод ранжування, Дельфі.*

Вступ

На сучасному етапі розвитку комп'ютерної графіки велику увагу приділяють формуванню високореалістичних зображень, для оцінювання якості яких використовують різні метрики.

Зокрема, оцінка якості рендерингу [1] графічних сцен здійснюється при використанні фізично-обґрунтованих (physically-based) моделей освітлення. Наприклад, при дослідженні фізично-точних двопробових функцій відбивної здатності (ДФВЗ) порівнюються синтетичні зображення з реальними фотографіями. Це дозволяє перевірити, наскільки точно ДФВЗ забезпечують відтворення оптичних властивостей матеріалів.

Окрім того, кожна з моделей освітлення, таких як Фонга, Кука-Торренса, Ламберта, Вальтера, має свої особливості застосування та обмеження. Оцінювання якості зображень дозволяє виявити невідповідності між змодельованими та реальними характеристиками відбиття світла від матеріалів. Відповідно, забезпечується підбір оптимальних параметрів освітлення та матеріалів для забезпечення більш реалістичної візуалізації сцени.

Оцінювання якості зображень також є важливим при використанні спрощених полігональних моделей об'єктів сцен і складних моделей освітлення, що використовуються у системах рендерингу реального часу. Спрощення геометричних полігональних моделей і використання апроксимаційних моделей освітлення може призводити до втрати важливих деталей об'єктів і відхилень у точності відтворення матеріалів. Метрики якості зображень дозволяють оцінити, наскільки спрощення забезпечують збереження візуальної інформації, і знайти компроміс між високою продуктивністю та фотореалістичністю.

Тому актуальним є питання оцінювання якості сформованих зображень.

Аналіз літератури

Для оцінювання якості зображень найчастіше використовуються аналітичні метрики. Аналітичні метрики якості розраховуються без залучення людини, згідно з математичними формулами на основі результатів обробки зображень. До основних аналітичних метрик якості [2] зображень належать *MSE*, *RMSE*, *NMSE*, *PSNR*, *SSIM*.

MSE [2] (Mean Square Error) полягає у знаходженні середнього квадратичного відхилення інтенсивностей кольору пікселів синтезованого та еталонного зображень. Використовується формула:

$$MSE = \left(\sum_{i=1}^I \sum_{j=1}^J (x_{ij} - y_{ij})^2 \right) / (IJ),$$

де I – кількість пікселів зображення по вертикалі, J – кількість пікселів зображення по горизонталі, x_{ij} – інтенсивність кольору пікселя еталонного зображення, y_{ij} – інтенсивність кольору пікселя тестового зображення.

Піднесення до квадрату у формулі застосовується для усунення від'ємних різниць між інтенсивностями кольору пікселів.

Чим менше значення *MSE*, тим більш якісне зображення.

RMSE [2] (Root Mean Square Error) полягає у обчисленні квадратного кореня значення *MSE*:

$$RMSE = \sqrt{MSE}.$$

Порівняно з *MSE*, через введення квадратного кореня забезпечується перехід до більш зрозумілої лінійної шкали похибок.

NMSE [3] (Normalized Mean Square Error), на відміну від *MSE*, включає нормування інтенсивностей кольору пікселів при обчисленні їх квадратичних різниць.

$$NMSE = \left(\sum_{i=1}^I \sum_{j=1}^J (x_{ij} - y_{ij})^2 \right) / \left(\sum_{i=1}^I \sum_{j=1}^J (x_{ij})^2 \right).$$

Зазвичай, якщо $NMSE > 0.001$, то зображення значно відрізняються. Якщо $NMSE < 0.0001$, візуальні відмінності між зображеннями є непомітними.

PSNR [2, 4] (Peak Signal to Noise Ratio) найчастіше використовується для оцінювання якості реконструйованих і стиснених зображень. Метрика визначається як відношення максимальної потужності сигналу до потужності шуму й розраховується згідно з виразом:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right).$$

Чим більше значення *PSNR*, тим якіснішим є зображення. Зокрема, під час тестування алгоритмів ущільнення зображень прийнятним є значення $PSNR > 33$ (дБ).

SSIM [2, 4] (Structural Similarity Index Measure) дозволяє розрахувати рівень структурної подібності між двома зображеннями, для чого обчислюються складові яскравості l , контрасту c і структури s .

Складова яскравості обчислюється згідно з формулою:

$$l = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1},$$

де μ_x – середнє значення інтенсивності кольору пікселя еталонного зображення, μ_y – середнє значення інтенсивності кольору пікселя тестового зображення, $c_1 = (0.01L)^2$, L – динамічний діапазон інтенсивності кольору пікселя.

Для обчислення складової контрасту використовується формула [2]:

$$c = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2},$$

де σ_x – стандартне відхилення інтенсивності кольору пікселя еталонного зображення, σ_y – стандартне відхилення інтенсивності кольору пікселя тестового зображення, $c_2 = (0.03L)^2$.

Структурна складова *SSIM* подає просторові взаємозалежності між пікселями та обчислюється згідно з виразом [2]:

$$s = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3},$$

де σ_{xy} – коваріація інтенсивностей пікселів еталонного та тестового зображень, $c_3 = c_2 / 2$.

Кінцева формула розрахунку *SSIM* [2] має вид:

$$SSIM = l^\alpha \cdot c^\beta \cdot s^\gamma,$$

де α, β, γ – підібрані степені (найчастіше 1).

SSIM характеризується діапазоном значень $[-1, 1]$, де 1 означає найвищу структурну подібність зображень.

Загалом, візуальна якість – суб'єктивна, і навіть найкращі аналітичні метрики не завжди узгоджуються з людським сприйняттям. Жодна аналітична метрика не є універсальною – різні метрики найкраще застосовувати при різних умовах (тестування алгоритмів рендерингу, ущільнення, згладження). Деякі метрики успішно застосовуються для виявлення лише певних типів спотворень зображень. Наприклад, *PSNR* неефективно застосувати для виявлення розмитих зображень. Також, більшість аналітичних метрик якості характеризуються використанням лише одного параметру зображення, наприклад, інтенсивності кольору пікселя.

Тому часто на практиці використовується експертне [5] (суб'єктивне) оцінювання якості зображень, яке, як правило, охоплює для аналізу кілька параметрів, які не завжди є корельовані.

Мета роботи – проаналізувати особливості основних методів і метрик експертного оцінювання якості зображень.

Аналіз особливостей експертного оцінювання якості зображень

Суб'єктивне оцінювання [6, 7] якості зображень, як правило, здійснюється групою експертів. Такий вид суб'єктивного оцінювання називається експертним.

Експертне оцінювання [5] тривимірних зображень – це процес, під час якого фахівці аналізують і оцінюють якість 3D-візуалізації. Це може включати оцінку таких аспектів, як точність моделювання поверхні, реалістичність текстур і матеріалів, а також загальна візуальна привабливість і відповідність зазначеним критеріям.

Експертне оцінювання зображень може застосовуватись у різних галузях.

У галузі комп'ютерних ігор експерти оцінюють якість рендерингу для забезпечення високого рівня реалістичності та відповідності візуальному стилю гри. Для цього аналізуються освітлення, текстури, відображення персонажів і середовища. Наприклад, при розробці ігор AAA-класу команди художників і технічних спеціалістів працюють разом, щоб досягти найвищої якості візуальних ефектів.

У кінематографі, особливо в жанрі фантастики та супергеройських фільмів, експерти з рендерингу оцінюють якість спецефектів і анімації для досягнення максимальної реалістичності та формування відповідного візуального враження. У фільмах використовуються передові технології рендерингу, що підлягають ретельній оцінці перед фінальним випуском. Зокрема, один із виразних прикладів залучення експертів до оцінювання якості рендерингу стосується виробництва анімаційного фільму «Людина-павук: Навколо Всесвіту», що відомий своїм унікальним стилем коміксів. Для розробки було

залучено команду висококваліфікованих експертів з анімації та візуальних ефектів, що працювали над ітеративним оцінюванням і вдосконаленням технік рендерингу. Експерти аналізували кожен сцену [8], щоб переконатися, що анімація відповідає високим стандартам якості, які вимагалися для відображення історії. Було зосереджено увагу на таких аспектах: підбір рівня текстурної деталізації і кольорової гами для передачі стилістичних елементів коміксів, моделювання тіней і освітлення для формування більш глибокого динамічного та візуального досвіду користувача, забезпечення оптимальної фреймової частоти для моделювання унікального «стрибокподібного руху», що імітує стиль коміксів. У результаті, фільм виграв нагороду «Оскар» за найкращий анімаційний фільм, частково завдяки якості рендерингу.

Архітектори та дизайнери застосовують 3D-візуалізацію для демонстрації проєктів клієнтам. Висока якість рендерингу дозволяє клієнтам краще уявити фінальний результат проєктування. Тому експерти з рендерингу оцінюють якість візуалізації, щоб забезпечити адекватне відтворення матеріалів і особливостей освітлення.

У галузі промислового дизайну експерти з рендерингу оцінюють 3D-моделі виробів, щоб впевнитися, що зображення точно передають дизайн, функціональність і матеріали, які будуть використані у реальному виробництві. Це особливо важливо для високотехнологічних виробів, таких як електроніка або автомобілі.

У галузі освіти експерти оцінюють якість візуалізації тривимірних моделей у навчальних програмах і програмах-симуляторах. Від рівня якості сформованих зображень залежить ефективність цих програм. Наприклад, у віртуальних хірургічних тренажерах, високоякісний рендеринг дозволяє лікарям краще зрозуміти анатомію та техніки хірургічних втручань.

У галузі наукових досліджень експерти можуть оцінювати якість візуалізації складних структур, як біологічних організмів або геологічних формацій.

Експертна перевірка якості зображень може здійснюватись візуально (найпростіший і найпопулярніший метод, перевіряються візуальні дефекти та недоліки зображень), на основі аналізу технічних параметрів (оптимальність кількості полігонів, використання технічних ресурсів), симуляційних тестів (результати використання моделі рендерингу в динамічному й інтерактивному середовищі), порівняльного аналізу з еталонними зображеннями, а також за допомогою автоматизованих інструментів оцінювання. Зокрема, під час експертного оцінювання можуть бути використані спеціальні плагіни таких 3D-засобів, як Autodesk Maya, Blender чи 3ds Max. Наприклад, плагіни включають функції для вимірювання рівня деталізації, густини полігональної моделі або характеристик використання ресурсів. Додатково, забезпечується перевірка зайвих вершин і непотрібних полігонів моделей, а також підтримка спільної роботи над проєктом рендерингу.

Експертні метрики та методи залежать від навичок, знань і психологічних особливостей експертів. Тому важливим є вдалий підбір експертів. Для забезпечення об'єктивності, компетентності та ефективності процесу оцінювання експерти повинні відповідати ряду критеріїв [9]:

1. Експерти повинні мати глибокі знання та досвід у відповідній галузі. Наприклад, при оцінюванні архітектурної візуалізації необхідно залучати архітекторів і дизайнерів інтер'єрів, а при оцінюванні кадрів відеоігор – професіоналів у галузі гейм-дизайну та комп'ютерної графіки.

2. Експерти мають мати досвід роботи зі схожими візуалізаційними проєктами. Такі експерти зможуть краще зрозуміти нюанси та вимоги завдання. Їхній практичний досвід може значно підвищити точність наданих оцінок.

3. Експерти повинні мати хорошу репутацію у власній галузі. Репутацію експерта можна перевірити шляхом аналізу відгуків, академічних праць, участі у конференціях, блогів і публічних виступів. Хороша репутація часто є свідченням надійності та професіоналізму експерта.

4. При виборі експертів особливо важливою є наявність високого рівня технічних навичок у галузях 3D-графіки та візуалізації. Експерти повинні володіти знаннями про використані програмне забезпечення та інструменти, щоб адекватно оцінити технічну якість сформованого зображення.

5. Додатково, експерти мають бути здатні об'єктивно оцінювати якість зображення, незалежно від особистих уподобань чи зв'язків з автором роботи.

6. Також, експерти повинні володіти високим рівнем комунікативних навичок. Експерти повинні вміти чітко та аргументовано викладати свої думки, щоб забезпечити зрозумілість та виконання їх рекомендацій.

7. Іншою важливою вимогою до експертів є гнучкість і адаптивність. Вимоги до якості зображень часто еволюціонують, тому експерти мають бути готові переглядати свої оцінки та відгуки в контексті нових даних і змін.

Важливим також є визначення оптимального числа експертів. Оптимальна кількість експертів для оцінювання може залежати від багатьох факторів, включаючи складність завдання, доступність кваліфікованих експертів, бюджет і часові рамки завдання. Зазвичай, рекомендується мати хоча б трьох експертів. Водночас, для більшості завдань краще мати від 5 до 7 експертів. Це число забезпечує достатню різноманітність думок, але все ще не вимагає значних координаційних зусиль. Для тестування складних методів формування зображень може бути залучена більша кількість експертів, щоб врахувати всі можливі погляди та підходи до системи. У таких випадках рекомендується ретельно структурувати процес оцінки.

Загалом, для експертного оцінювання якості зображень можуть бути застосовані два основні типи метрик (з одним стимулом і двома стимулами) та два типи експертних методів (анкетування й групової експертизи).

Аналіз метрик експертного оцінювання якості зображень

Обчислення метрик якості з одним стимулом полягає у оцінюванні експертами якості лише тестових зображень за допомогою визначеної шкали та усередненні результатів.

Поширеною є метрика *MOS* (Mean Opinion Score) [10]. Значення метрики отримується у результаті методу присвоєння абсолютних категорій якості (Absolute Category Ranking). Кожен експерт обирає рівень якості зображення з переліку: «погана», «низька», «середня», «добра», «відмінна». Ці рівні якості зображень відповідають числовим оцінкам діапазону [1,5]. Числові відповідники оцінок додаються та усереднюються. Використовується формула:

$$MOS = \frac{1}{m} \sum_{j=1}^m s(j),$$

де $s(j)$ – оцінка рівня якості тестового зображення, m – кількість експертів.

Шкала *MOS* може бути застосована як і для загального оцінювання якості зображення, так і для оцінювання окремих компонент якості. Наприклад, експерти можуть оцінювати реалізм освітлення, деталізацію текстур, час рендерингу.

У таблиці 1 наведено гіпотетичні відповідності висновків експерта щодо якості зображення та оцінок шкали *MOS*.

Приклад застосування шкали *MOS* для оцінювання аспектів якості 3D-зображення

Аспект якості	Виставлена експертом оцінка якості	Суб'єктивний висновок експерта щодо якості зображення
Реалізм освітлення	1 – Погана	Освітлення нехарактерне для джерела світла.
	2 – Низька	Моделювання освітлення характеризується помітними помилками.
	3 - Середня	Освітлення в основному змодельовано правильно, хоча деякі аспекти можуть бути покращені.
	4 – Добра	Освітлення здебільшого реалістичне.
	5 – Відмінна	Освітлення відповідає вищим стандартам фотореалізму.
Деталізація Текстур	1 – Погана	Текстури виглядають розмиті або нечітко, без деталей.
	2 – Низька	Текстури мають низьку деталізацію, помітні артефакти.
	3 - Середня	Текстури відповідають задуму художника, але потребують покращення.
	4 – Добра	Текстури високої якості з хорошою деталізацією.
	5 – Відмінна	Текстури дуже деталізовані та реалістичні.
Час рендерингу	1 – Погана	Час рендерингу значно вищий за прийнятні норми.
	2 – Низька	Час рендерингу вищий, ніж бажано для наявного рівня якості.
	3 – Середня	Час рендерингу адекватний, але може бути вдосконалений.
	4 – Добра	Час рендерингу майже оптимальний і відповідає якості сцени.
	5 – Відмінна	Час рендерингу оптимальний, забезпечує високу продуктивність.

Метрика є простою, однак потребує підбору порогового значення класифікації зображення як якісного.

Іншою метрикою є *SSCQE* (Single-Stimulus Continuous Quality Evaluation) [7]. Використовується неперервна шкала значень якості зображення. Обраний експертом рівень якості на неперервній шкалі відповідає значенню числовому діапазону [0, 100]. Далі, обчислюється середнє значення метрики.

Розрахунок метрик якості з двома стимулами передбачає оцінювання експертами якості еталонного та тестового зображень.

Метрика *DSIS* (Double Stimulus Impairment Scale) [7, 10] використовується для визначення пошкоджень тестового зображення порівняно з еталонним. Як правило, експертам по чергові показують еталонне та тестове зображення, розділені кадром із сірим фоном. Експерти визначають рівень пошкоджень тестового зображення як «дуже докучливий», «докучливий», «дещо докучливий», «відчутний, але не докучливий», «непомітний». Ці рівні, як і у випадку метрики *MOS*, відповідають числовим значенням діапазону [1, 5]. Після опитування експертів визначається середнє значення *DSIS*.

Метрика *DMOS* (Difference Mean Opinion Score) [6, 11], на відміну від *MOS*, полягає в усередненні різниць експертних оцінок якості тестового та еталонного зображень. *DMOS* обчислюється згідно з формулою:

$$DMOS = \frac{1}{m} \sum_{j=1}^m rs(j) - s(j),$$

де $rs(j)$ – рівень якості еталонного зображення.

DSCQS (Double-Stimulus Continuous Quality Score) [7], на відміну від *SSCQE*, передбачає оцінювання якості тестового та еталонного зображень за допомогою неперервної шкали. Причому, експертам не повідомляється, яке із зображень є еталонним. У подальшому, знаходиться середнє відхилення оцінок якості зображень.

Аналіз методів експертного оцінювання якості зображень

Методи анкетування полягають у опитуванні групи експертів щодо якості зображення за допомогою анкет й у подальшому опрацюванні результатів опитування. Методи анкетування є відносно дешевими, дозволяють одночасно з'ясувати думку значної кількості експертів. Однак, експерти можуть неправильно зрозуміти окремі запитання або поставитись до них несерйозно.

Метод ранжування [5] може застосовуватись, насамперед, для оцінювання значимості факторів певної характеристики. Для оцінювання якості зображень факторами можуть бути: точність відтворення відблисків, точність формування тіней, відсутність артефактів, рівень контрастності, чіткість контурів об'єктів. Метод ранжування полягає у тому, що кожен експерт присвоює ранг значимості фактору якості зображення (ранг 1 відповідає найбільшій значимості). При цьому, для забезпечення чіткого розрізнення факторів, їх число повинне бути менше 20. У результаті опитування групи експертів формується спеціальна матриця опитування, де рядки відповідають рангам одного експерта для різних факторів, а стовпці відповідають рангам різних експертів для одного фактору. Далі, формується матриця перетворених рангів, де у кожному стовпці для фактора обчислюється сума рангів R_i . Ранги у комірках таблиці змінюються шляхом віднімання від найбільшого рангу виставлених експертами значень. Після цього, обчислюються відносні ваги кожного фактору W_i . Відносна вага [12] для окремого фактору знаходиться шляхом ділення суми його рангів на суму сум рангів усіх факторів:

$$W_i = R_i / \sum_{i=1}^n R_i$$

У результаті, отримуються коефіцієнти різних факторів для зваженого показника рівня якості зображення. Зважений показник якості зображення обчислюється згідно з виразом:

$$\sum_{i=1}^n W_i \cdot F_i,$$

де W_i – вага параметра, F_i – нормалізоване значення параметра.

Важливим є встановлення рівня узгодженості експертних оцінок рангів при анкетуванні. Для цього використовується коефіцієнт конкордації Кендалла [13]

$$V = \frac{12S}{m^2(n^3 - n)},$$

де m – кількість експертів, n – кількість факторів якості зображення, S – сума квадратичних відхилень сум рангів окремих факторів від середньої суми рангів факторів,

$$S = \sum_{i=1}^n (R_i - \bar{R}_i)^2.$$

Діапазон значень коефіцієнта становить [0, 1]. Значення коефіцієнта конкордації Кендалла, більше за 0.4, означає середню узгодженість експертів. Висока узгодженість [5] експертів досягається за значення коефіцієнта, більшого за 0.6–0.8.

Загалом, обчислення коефіцієнта Кендалла дозволяє визначити, наскільки надійні та узгоджені оцінки рангів, що даються різними експертами, дозволяючи визначити аспекти, щодо яких існує значна неоднозначність або розбіжності.

Окрім того, цей метод анкетування може бути застосований для ранжування групи зображень за суб'єктивним рівнем якості. У такому випадку, кожен експерт присвоює ранги зображенням з набору відповідно до їх рейтингу якості. При цьому, декільком зображенням може бути присвоєний однаковий ранг. Наприкінці, для кожного зображення обчислюється середній ранг (рейтинг). Найбільш якісним є зображення у наборі із найменшим середнім рангом.

Розглянемо приклад застосування методу ранжування (табл. 2). Вважатимемо, що для п'яти характеристик зваженого показника якості візуалізації необхідно обчислити відносні ваги. Троє експертів упорядковують характеристики за рівнем значимості. Кожній характеристиці присвоюється ранг, де 1 відповідає найбільшій значимості, а 5 – найменшій.

Таблиця 2

Приклад застосування методу ранжування для упорядкування характеристик якості візуалізації 3D-сцен

Характеристика	Продуктивність рендерингу (A)	Точність відтворення відблисків (B)	Точність відтворення тіней (C)	Рівень деталізації текстур (D)	Чіткість контурів (E)
Експерт					
Експерт 1	1	2	3	5	4
Експерт 2	1	2	4	3	5
Експерт 3	2	1	3	4	5

Переходимо до матриці перетворених рангів (таблиця 3) – у кожній клітинці таблиці 2 віднімаємо від максимального рангу значення виставленого експертом рангу.

Таблиця 3

Матриця перетворених рангів

Характеристика	Продуктивність Рендерингу (A)	Точність відтворення відблисків (B)	Точність відтворення тіней (C)	Рівень деталізації текстур (D)	Конструктивна точність (E)
Експерт					
Експерт 1	4	3	2	0	1
Експерт 2	4	3	1	2	0
Експерт 3	3	4	2	1	0

Обчислюємо суму перетворених рангів R_i кожної характеристики якості зображення:

$$R_i(A) = 4 + 4 + 3 = 11; \quad R_i(B) = 3 + 3 + 4 = 10; \quad R_i(C) = 2 + 1 + 2 = 5;$$

$$R_i(D) = 0 + 2 + 1 = 3; \quad R_i(E) = 1 + 0 + 0 = 1.$$

Шляхом ділення R_i на суму усіх R_i знаходимо вагові коефіцієнти W_i характеристик якості зображення, що можуть бути використані при обчисленні зваженого показника якості:

$$W_i(A) = 11 / (11 + 10 + 5 + 3 + 1) = 0.367; \quad W_i(B) = 10 / 30 = 0.333; \quad W_i(C) = 5 / 30 = 0.167;$$

$$W_i(D) = 3 / 30 = 0.100; \quad W_i(E) = 1 / 30 = 0.033.$$

Встановимо рівень узгодженості присвоєння експертних рангів. Обчислюємо середнє значення сум рангів \bar{R}_i :

$$\bar{R}_i = (11 + 10 + 5 + 3 + 1) / 5 = 6.$$

Визначимо суму квадратів відхилень S значень R_i від \bar{R}_i :

$$S = (11 - 6)^2 + (10 - 6)^2 + (5 - 6)^2 + (3 - 6)^2 + (1 - 6)^2 = 76.$$

Обчислюємо коефіцієнт конкордації Кендалла V :

$$V = \frac{12S}{m^2(n^3 - n)} = \frac{12 \cdot 76}{3^2(5^3 - 5)} = 0.84.$$

Значення $V = 0.84$ є високим й означає достатньо сильну узгодженість думок експертів.

Іншим методом анкетування для експертного оцінювання якості зображень є метод нормування. Метод полягає у тому, що експерт ставить у відповідність кожному фактору якості зображення певний рівень шкали значимості $[0, 1]$. У результаті анкетування групи

експертів формується матриця опитування. Аналогічно до методу ранжування, рядки містять виставлені експертами ваги для різних факторів, а стовпці містять ваги, виставлені різними експертами для одного фактору. Спершу, знаходяться суми ваг B_j , виставлені кожним j -им експертом. Далі, здійснюється нормування – вага w_{ij} у кожній комірці матриці ділиться на B_j відповідного рядка матриці. Відносна вага окремого фактору W_i [14] визначається як сума значень w_{ij} у стовпці i -го фактору, поділена на суму усіх значень w_{ij} .

$$W_i = \left(\sum_{j=1}^m w_{ij} \right) / \left(\sum_{j=1}^m \sum_{i=1}^n w_{ij} \right).$$

Метод нормування, як і метод ранжування, дозволяє визначити ваги факторів для обчислення зваженого показника якості зображення.

Для порівняння якості зображень набору та визначення ваг характеристик якості зображення може бути використаний метод попарного порівняння [15]. Метод полягає у формуванні матриці, де записуються результати попарного порівняння зображень. Наприклад, краща якість при порівнянні може позначатись одиницею, гірша якість – нулем, елементи головної діагоналі – знаком «-». Після опитування усіх експертів формується кінцева матриця, комірки якої містять сумарні результати порівнянь. Відповідно, зображення із найкращою якістю характеризується найбільшою сумою елементів відповідного рядка.

Аналогічно, метод порівнянь застосовується при визначенні ваг характеристик якості зображень. Використовується вираз [15]:

$$W_i = \sum_{j=1}^m \frac{f_{ij}}{C},$$

де C – кількість можливих порівнянь, f_{ij} – частота надання переваги j -им експертом i -ій властивості над іншими властивостями.

Недоліками методу є те, що кількість порівнянь може бути великою й на рішення експерта можуть впливати психологічні фактори, як втома.

Окрім розглянутих методів, існують методи анкетування, що базуються на застосуванні відомих шкалах. Наприклад, для оцінювання якості зображень може використовуватись шкала Лайкерта [16].

Шкала Лайкерта часто використовується для оцінки відповідей у соціологічних і психологічних дослідженнях, однак також придатна для оцінювання якості візуалізації 3D-сцен. Шкала має зазвичай 5 рівнів відповідей на запитання, що дозволяють оцінити аспект якості тривимірного зображення: «повністю не згоден», «не згоден», «десь посередині», «згоден», «повністю згоден».

При оцінюванні якості зображень запитання можуть бути пов'язані з такими аспектами якості, як реалізм освітлення (відтворення відблисків і тіней), деталізація текстур, візуальна привабливість (загальний вигляд, кольорова гама), час рендерингу.

Розглянемо приклад застосування шкали Лайкерта для оцінювання рендерингу тривимірної сцени (табл. 4). Таблиця містить твердження щодо різних аспектів якості зображення, а також відповідні відповіді одного експерта.

Таблиця 4

Застосування шкали Лайкерта для оцінювання якості 3D-зображення

Аспект Якості	Запитання	Повністю не згоден	Не згоден	Десь посередині	Згоден	Повністю згоден
Реалізм освітлення	Модель освітлення відтворює реальні умови освітленості інтер'єру				+	
Деталізація Текстур	Текстури на меблях та на стінах виглядають високодеталізовано та реалістично					+
Загальна візуальна привабливість	Загальний візуальний вигляд моделі є привабливим і естетично приємним				+	
Реалізм відображення матеріалів	Матеріали, як-от скло, метал і тканина, відображені з високим рівнем реалізму				+	
Час рендерингу	Час, необхідний для рендерингу цієї моделі, є прийнятним			+		

Після збору відповідей експертів, наданих за допомогою шкали Лайкерта, здійснюється систематизація та обробка отриманих даних.

Обробка даних передбачає розрахунок статистичних показників, таких як середнє значення, мода, медіана. Наприклад значення шкали «повністю не згоден» - «згоден» переводяться у числову шкалу діапазону [1,5]. Далі, знаходяться середні виставлені оцінки по кожному запитанню, кожному аспекту якості й загалом. У такому випадку, отримане середнє значення є аналогічним значенню метрики MOS за винятком того, що використовується інша словесна шкала якості.

Для оцінки узгодженості експертних бальних оцінок якості зображення обчислюється значення стандартного відхилення. Стандартне відхилення розраховується як корінь усередненої суми квадратичних відхилень елементів вибірки від їх середнього значення.

Наприклад, у таблиці 5 наведено переведені у числову шкалу оцінки 4 експертів для 4 зображень на основі шкали Лайкерта.

Таблиця 5

Приклад виставлених зображенням балів на основі шкали Лайкерта

Зображення	Зображення 1	Зображення 2	Зображення 3	Зображення 4
Експерти				
Експерт 1	5	3	5	5
Експерт 2	4	2	3	5
Експерт 3	3	3	3	5
Експерт 4	5	4	4	4

Обчислимо середні значення оцінок $\mu(i)$ для кожного зображення:

$$\mu(1) = (5 + 4 + 3 + 5) / 4 = 4.25, \quad \mu(2) = (3 + 2 + 3 + 4) / 4 = 3,$$

$$\mu(3) = (5 + 3 + 3 + 4) / 4 = 3.75, \quad \mu(4) = (5 + 5 + 5 + 4) / 4 = 4.75.$$

На основі значень $\mu(i)$ обчислюємо значення стандартного відхилення $\sigma(i)$ для оцінок усіх зображень.

$$\sigma(1) = \sqrt{\frac{(5-4.25)^2 + (4-4.25)^2 + (3-4.25)^2 + (5-4.25)^2}{4}} = 0.83, \quad \sigma(2) = \sqrt{\frac{(3-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2}{4}} = 0.71,$$

$$\sigma(3) = \sqrt{\frac{(5-3.75)^2 + (3-3.75)^2 + (3-3.75)^2 + (4-3.75)^2}{4}} = 0.83, \quad \sigma(4) = \sqrt{\frac{(5-4.75)^2 + (5-4.75)^2 + (5-4.75)^2 + (4-4.75)^2}{4}} = 0.43.$$

Чим менше значення $\sigma(i)$, тим більш узгоджені оцінки окремого зображення. У цьому випадку, найбільша узгодженість експертів наявна при оцінюванні четвертого зображення.

Більш складний метод дисперсійного аналізу ANOVA [17] дозволяє виявити статистично значущі відмінності (не пояснюються простими коливаннями значень) між групами оцінок якості зображень. Наприклад, перевіряється, чи відрізняється суттєво якість зображень набору.

Розглянемо приклад дисперсійного аналізу на основі таблиці 3. Знаходимо середнє значення $\mu(i)$:

$$\mu = (4.25 + 3 + 3.75 + 4.75) / 4 = 3.94.$$

Обчислюємо міжгрупову варіативність (пояснену дисперсію) згідно з виразом:

$$\frac{\sum_{i=1}^K n_i (\mu(i) - \mu)^2}{K - 1} = \frac{4(4.25 - 3.94)^2 + 4(3 - 3.94)^2 + 4(3.75 - 3.94)^2 + 4(4.75 - 3.94)^2}{4 - 1} = 2.23.$$

K позначає кількість груп (у цьому випадку – кількість зображень), n_i – кількість спостережень (оцінок) групи.

Розраховуємо значення внутрішньогрупової варіативності (непоясненої дисперсії) згідно з формулою:

$$\frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (Y_{ij} - \mu(i))^2}{N - K} = \frac{(5-4.25)^2 + (4-4.25)^2 + (3-4.25)^2 + (5-4.25)^2 + (3-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3.75)^2 + (3-3.75)^2 + (3-3.75)^2 + (4-3.75)^2 + (5-4.75)^2 + (5-4.75)^2 + (5-4.75)^2 + (4-4.75)^2}{16 - 4} = 0.69.$$

N є сумарним розміром вибірки.

Далі, шляхом ділення поясненої дисперсії на непояснену розраховується показник F .

$$F = 2.23 / 0.69 = 3.23$$

F порівнюється зі спеціальним критичним значенням F_k , що береться з таблиці на основі рівня значущості (наприклад, 5%), чисел K і N . Якщо $F > F_k$, відмінності між групами є статично значущими. У цьому випадку, $F_k = 3.49$. Отже, хоч і середні оцінки зображень є достатньо різними, статистичні відмінності між оцінками якості різних зображень не є значущими. Відповідно, якість зображень набору вважається схожою. Це пояснюється тим, що оцінки експертів є дещо розкиданими, обраним рівнем статистичної значущості, а також невеликою кількістю зображень.

Аналогічно, можливо перевірити статистичні відмінностями між наборами (групами) оцінок експертів.

В подальшому, здійснюється візуалізація результатів аналізу даних, формування звітів з рекомендаціями, внесення необхідних змін до алгоритмів рендерингу, здійснюється моніторинг результатів внесення змін.

Застосування методу семантичного диференціалу [18] для оцінювання якості зображення полягає у використанні набору шкал, що використовують пари протилежних прикметників. Наприклад, використовуються шкали «Нереалістична сцена» – «Фотореалістична сцена», «Наявні видимі артефакти» – «Відсутні видимі артефакти», «Деталі об'єктів розмиті» – «Деталі об'єктів чіткі». Зазвичай окрема шкала семантичного диференціалу включає п'ять або сім рівнів. Проміжні рівні шкали можуть бути подані як числа (наприклад, [1,5] чи [-2,2]) або слова, такі як «зовсім не», «досить», «дуже». Після збору даних здійснюється їх

статистичне опрацювання, зокрема, знаходяться середні оцінки якості для кожної шкали.

Методи групової експертизи дозволяють сформувавши групову експертну думку щодо якості зображення. Найбільш відомим методом групової експертизи є метод Дельфі [19].

На кожному етапі методу здійснюється анонімне опитування експертів щодо якості зображення. Анонімність опитування забезпечує можливість висловлення думки усіма експертами. Далі, здійснюється статистичний аналіз відповідей експертів для формування групової відповіді. Як правило, визначаються медіанна (серединна) оцінка якості зображення та квартилі оцінок якості. На проміжних етапах методу експерти отримують результати статичного аналізу. Відповідно, експерти, оцінки яких потрапили у верхній і нижній квартилі (25 % найменших і 25 % найвищих оцінок), мають можливість переглянути власні оцінки. У результаті, групова оцінка якості стає все більш однорідною. Зазвичай, всього застосовується 3 – 4 раунди.

Загальні висновки та консенсус, досягнуті на останньому етапі опитування, компілюються у звіт, який містить рекомендації для вдосконалення використаних алгоритмів рендерингу. Метод дозволяє зібрати глибокі та обґрунтовані оцінки експертів, мінімізуючи вплив домінування окремих особистостей чи упереджень.

Іншими методами групової експертизи [20,21] є метод консенсусної панелі, метод номінальних груп, метод снігової кулі, експертні інтерв'ю. Метод консенсусної панелі полягає у організації відкритої панельної дискусії між експертами. Кінцевий висновок базується на основі досягнутого консенсусу. Метод номінальних груп, на відміну від методу консенсусної панелі, характеризується більш структурованим обговоренням ідей. Спершу, ідеї індивідуально записуються експертами, а потім представляються для групового обговорення та голосування. Характерною рисою методу снігової кулі є початкова незначна кількість експертів, які в подальшому рекомендують інших фахівців до участі в оцінюванні. Метод експертних інтерв'ю дозволяє глибше зануритися в специфічні аспекти оцінювання якості зображення. Метод полягає у здійсненні глибинних індивідуальних інтерв'ю з експертами.

Зрозуміло, що методи групової експертизи доцільно застосовувати лише у випадку тестування складних алгоритмів формування зображень, коли якість візуалізації є особливо важливою.

Висновки

Експертні метрики та методи забезпечують оцінювання якості зображень із врахуванням особливостей людського сприйняття. Метрики з одним стимулом використовуються для розрахунку середнього рівня якості тестового зображення на основі оцінок експертів. Метрики з двома стимулами призначені для оцінювання експертами середнього відхилення у рівні якості між тестовим і еталонним зображеннями. Експертні методи анкетування дозволяють здійснити сортування набору зображень за рівнем якості, а також розрахувати коефіцієнти зваженого критерію якості. Методи групової експертизи використовуються для формування групової оцінки якості зображення у специфічних ситуаціях.

СПИСОК ЛІТЕРАТУРИ

1. Романюк О. Н., Романюк О. В., Чехмestрук Р. Ю. Комп'ютерна графіка. Вінниця : ВНТУ, 2023. 146 с.
2. Завальнюк Є. К., Романюк О. Н. Огляд метрик порівняння якості зображень. *Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2023)», м. Вінниця, 22 трав. 2023 р. Вінниця, 2023. С. 571–573.*
3. Romanyuk O., Zavalniuk Y. Deep Learning-Based Determination of Optimal Triangles Number of Graphic Objects Polygonal Model. *IntelITSIS'2024: 5th International Workshop on Intelligent Information Technologies and Systems of Information Security*, Khmelnytskyi, 28 March 2024. Khmelnytskyi. P. 39–51.
4. Iesalnieks J. Full-Reference Quality Metrics: VMAF, PSNR and SSIM. *TestDevLab.com*. URL: <https://www.testdevlab.com/blog/full-reference-quality-metrics-vmf-psnr-and-ssim> (date of access: 28.02.2025).
5. Основні критерії оцінювання продуктивності та реалістичності систем кінцевої візуалізації /

- О. Н. Романюк та ін. *Інформаційні технології і автоматизація - 2021*, м. Одеса, 21–22 жовт. 2021 р. Одеса, 2021. С. 72–74.
6. The analysis of subjective metrics and expert methods for image quality assessment / O. N. Romanyuk et al. *Intellectual capital is the foundation of innovative development '2023*. Karlsruhe, 2023. P. 131–138.
7. Testolina M., Ebrahimi T. Review of subjective quality assessment methodologies and standards for compressed images evaluation. *Applications of Digital Image Processing XLIV*, San Diego, United States, 1–5 August 2021 / ed. by A. G. Tescher, T. Ebrahimi. 2021. URL: <https://doi.org/10.1117/12.2597813> (date of access: 14.02.2025).
8. Pearson B. Spider-Man: Into The Spider-Verse' Directors Explain Creating The Film's Unique Look. *SlashFilm*. URL: <https://www.slashfilm.com/562696/spider-man-into-the-spider-verse-interview/> (date of access: 28.02.2025).
9. Баланчук І. С., Бессараб В. Г. Розроблення критеріїв відбору експертів для проведення наукової та науково-технічної експертизи проєктів: аналіз світового досвіду. *Наука, технології, інновації*. 2021. № 4. С. 49–59.
10. Sugito Y., Bertalmio M. Non-Experts or Experts? Statistical Analyses of MOS using DSIS method. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, 4–8 May 2020. P. 2732–2736.
11. Non-Experts or Experts? Statistical Analyses of MOS using DSIS method. / T. P. Thanh et al. *International Conference on Knowledge and Systems Engineering (KSE)*, Can Tho, 12–14 November 2020. P. 55–60.
12. Intellectual potential assessing methodology of an innovation-oriented enterprise / H. Y. Ostrovska et al. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*. 2024. № 4. P. 141–148. URL: <https://doi.org/10.33271/nvngu/2024-4/141> (date of access: 15.02.2025).
13. Franceschini F., Maisano D. Decision concordance with incomplete expert rankings in manufacturing applications. *Research in Engineering Design*. 2020. Vol. 31. P. 471–490.
14. Стеблюк Н. Ф., Опаренко В. В. Визначення ефективності управління інвестиційною діяльністю підприємства з використанням методу «Делфі». *Інфраструктура ринку*. 2018. Вип. 25. С. 470–477.
15. Групове експертне оцінювання та компетентність експертів / О. М. Величко та ін. Одеса: ФОП Бондаренко М. О., 2015. 286 с.
16. How subjective CT image quality assessment becomes surprisingly reliable: pairwise comparisons instead of Likert scale / E. J. Hooijmakers et al. *European radiology*. 2024. Vol. 4. P. 4494–4503.
17. Siegel A. F., Wagner M. R. Practical business statistics. Cambridge: Academic press, 2021. 560 p.
18. Стамат В. М. Маркетинг. Миколаїв: МНАУ, 2024. 172 с.
19. Міца О. В., Лавер В. О. Системний аналіз. Ужгород: ПП «АУТДОР-ШАРК», 2021. 63 с.
20. Федоров О. В., Цумарев М. І., Лебедев О. П. Використання активних методів навчання психологічних, юридичних, економічних дисциплін у бізнес-тренінгах. *Соціально-правові стадії*. 2021. Вип. 4 № 14. С. 184–189.
21. Луцяк В. В., Остапчук А. Д., Бондарчук Я. П. Особливості управління практикою дослідження зв'язків з громадкістю. *Економіка АПК*. 2020. № 12. С. 89–102.

Стаття надійшла до редакції 04.03.2025.

Стаття пройшла рецензування 09.03.2025.

Завальнюк Євген Костянтинович – аспірант кафедри програмного забезпечення.

Романюк Олександр Никифорович – д-р техн. наук, професор, завідувач кафедри програмного забезпечення, e-mail: rom8591@gmail.com.

Вінницький національний технічний університет.

Ковальова Юлія Олександрівна – інженер.

Бобко Олексій Леонідович – аспірант кафедри програмного забезпечення.

Вінницький національний технічний університет.