KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

MIARA PODOBIEŃSTWA STRUKTURALNEGO DLA WIZJI WSZECHOGARNIAJĄCEJ

STRUCTURAL SIMILARITY METRIC FOR IMMERSIVE VIDEO

Jakub Stankowski, Weronika Nowak, Tomasz Grajek, Adrian Dziembowski

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{jakub.stankowski, tomasz.grajek, adrian.dziembowski}@put.poznan.pl

Streszczenie: **W pracy przedstawiono nową miarę obiektywnego pomiaru jakości dla wizji wszechogarniającej. Zaproponowana metryka powstała w sposób analogiczny do tego jak opracowano miarę IV-PSNR na podstawie miary PSNR. Tym razem zamiast PSNR wykorzystano SSIM. Przeprowadzone eksperymenty pokazały, że zaproponowana miara SSIV jest jeszcze lepiej skorelowana z wynikami oceny subiektywnej (MOS) niż inne znane metryki, w tym IV-PSNR.**

Słowa kluczowe: **wizja wszechogarniająca, ocena jakości wizji, obiektywny pomiar jakości.**

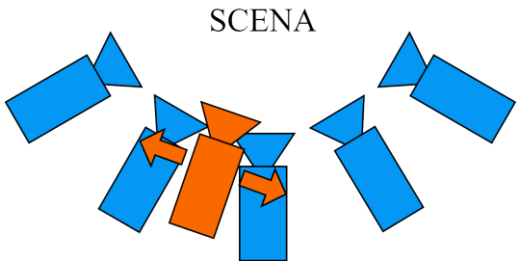
Abstract: **In this paper, a new objective quality metric for immersive video is presented. The proposed metric was created in a manner analogous to how the IV-PSNR metric was based on the PSNR one. This time, SSIM was used instead of PSNR. The conducted experiments showed that the proposed SSIV metric is even higher correlated with the results of subjective evaluation (MOS) than other known metrics, including IV-PSNR.**

Keywords: **immersive video, video quality evaluation, objective quality assessment.**

1. WSTĘP

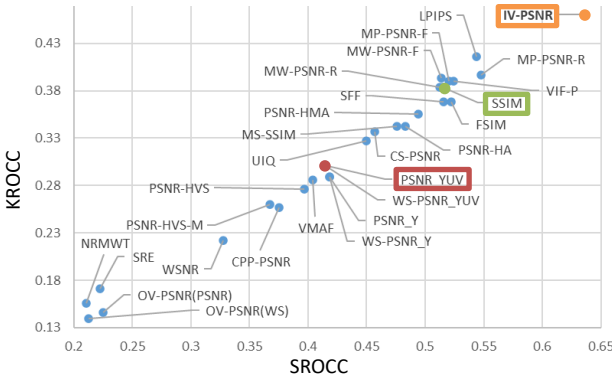
Rozwój technologiczny wpływa na wiele dziedzin naszego życia. W systemach wizyjnych można zauważyć coraz bardziej skomplikowane rozwiązania pozwalające w coraz lepszym stopniu odwzorowywać otaczającą nas rzeczywistość. Najnowsze rozwiązania dotyczą tzw. wizji wszechogarniającej (ang. *Immersive vision*) [1]. Systemy takie pozwalają użytkownikowi/widzowi „zanurzyć się" w prezentowanej treści i dowolnie zmieniać położenie, z którego obserwuje się scenę [2], [4]. Ideę takiego systemu przedstawiono na rysunku 1, gdzie niebieskie kamery oznaczają rzeczywiste kamery, którymi zarejestrowano daną scenę, natomiast kamera pomarańczowa oznacza użytkownika, który może zmieniać punkt obserwacji sceny. Oczywistym jest, że przy takim podejściu, dla większości położeń kamery pomarańczowej nie mamy danych wizyjnych dostępnych bezpośrednio z rzeczywistych kamer, więc dane te należy wytworzyć (wyrenderować) wykorzystując np. algorytm syntezy widoków wirtualnych [2], [4].

Skomplikowanie systemów wizyjnych wraz z nowymi sposobami generowania i prezentowania danych wizyjnych prowadzi do powstawania nowych typów zniekształceń wynikających z przetwarzania wizji wszechogarniakącej. W związku z tym, rzetelna ocena jakości danych wizyjnych staje się coraz trudniejsza.



Rys. 1. Idea systemu wizji wszechogarniającej; niebieskie – kamery rzeczywiste (wejściowe), pomarańczowa – kamera wirtualna.

Aby uwzględnić zniekształcenia typowe dla wizji immersyjnej, wprowadzono metrykę IV-PSNR [3], która uwzględnia typowe dla niej artefakty spowodowane przerzutowaniem informacji między widokami, a więc niewielkie przesunięcie obiektów i globalną zmianę charakterystyki barwnej renderowanego (syntezowanego) widoku wirtualnego.



Rys. 2. Korelacja (mierzona z użyciem metryk SROCC i KROCC [7]) pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie, z użyciem metryk opisanych w literaturze. Rysunek z [3].

Jak przedstawiono na rys. 2, dostosowana do charakterystyki wizji wszechogarniającej metryka IV-PSNR, wykazuje wyższą korelację z oceną jakości subiektywnej niż inne metryki (wliczając w to także metryki dedykowane dla obrazów syntezowanych, takie jak MP-PSNR-F [9], MP-PSNR-R [11] oraz MW-PSNR-F [10]). Jednakże,

metryka ta bazuje na PSNR (kolor czerwony, rys. 2), który jest zdecydowanie mniej efektywny niż pomiar podobieństwa strukturalnego (SSIM – kolor zielony, rys. 2).

W niniejszej pracy przedstawiono metrykę łączącą zalety obu technik, dopasowując pomiar podobieństwa strukturalnego do potrzeb wizji wszechogarniającej w sposób analogiczny jak w przypadku metryki IV-PSNR.

2. WSKAŹNIK PODOBIEŃSTWA STRUKTURALNEGO (SSIM)

Wskaźnik podobieństwa strukturalnego – SSIM (*Structural Similarity Index Measure*) [15] jest obiektywną miarą jakości, która od czasu publikacji zyskała bardzo dużą popularność w zastosowaniach związanych z oceną jakości obrazu. Jej implementacje można znaleźć w popularnych narzędziach takich jak MATLAB, bibliotekach programistycznych (np. scikit-image), czy nawet pakietach do uczenia maszynowego (np. TensorFlow). Wiele popularnych implementacji koderów wizyjnych, takich jak x264 [17] (koder AVC [13]), x265 [18] (koder HEVC [12]) czy libaom [16] (koder AV1 [5]) pozwala na użycie SSIM jako miary jakości w procesie optymalizacji przepływność-zniekształcenia RD (*rate – distortion*).

Miara SSIM obliczana jest jako iloczyn ważony, składający się z trzech składników zdefiniowanych jako: luminancja[1] ($\boldsymbol{L}$), kontrast ($\boldsymbol{L}$) i struktura ($\boldsymbol{S}$). Wspomniane składniki wyliczane są na bazie lokalnych statystyk takich jak: wartość średnia, odchylenie standardowe i kowariancja.

Cechą specyficzną dla wskaźnika podobieństwa strukturalnego jest wyliczanie powyższych statystyk w oknie o rozmiarze 11x11, gdzie wartości próbek wejściowych są ważone dwuwymiarową maską Gaussowską $\boldsymbol{\omega}$ (o odchyleniu standardowym równym 1,5) [14].

3. PODOBIEŃSTWO STRUKTURALNE W OCENIE WIZJI WSZECHOGARNIAJĄCEJ (SSIV)

W niniejszym artykule zaproponowano metrykę stworzoną do oceny jakości wizji wszechogarniającej. Zaproponowana metryka jest oparta na oryginalnej koncepcji wskaźnika podobieństwa strukturalnego ale została przeprojektowana w taki sposób aby dopasować ją do opisanych w [3] zniekształceń.

Pierwszym etapem wyznaczania podobieństwa strukturalnego obrazów I i J jest wyznaczenie lokalnych statystyk tychże obrazów: średniej ($\boldsymbol{\mu}^I$, $\boldsymbol{\mu}^J$), odchylenia standardowego ($\boldsymbol{\sigma}^I$, $\boldsymbol{\sigma}^J$) oraz kowariancji ($\boldsymbol{\sigma}^{I,J}$) dla każdej składowej barwnej c obu obrazów.

Dla każdego piksela (x, y) analizowane jest otoczenie $[2k+1] \times [2k+1]$, a wartości pikseli z tegoż otoczenia ważone są z użyciem maski $\boldsymbol{\omega}$. W przypadku obrazu I średnia i odchylenie standardowe wyznaczane są w taki sam sposób, jak w metryce SSIM [15]:

$$\boldsymbol{\mu}_c^I(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{I}_c(i,j)] \,, \tag{1}$$

$$\boldsymbol{\sigma}_c^I(x,y) = \sqrt{\sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\boldsymbol{\omega}(i-x, j-y) \cdot \left(\boldsymbol{I}_c(i,j)\right)^2\right] - \left(\boldsymbol{\mu}_c^I(x,y)\right)^2} \,, \tag{2}$$

Pozostałe lokalne statystyki wyznaczane są w sposób zmodyfikowany, dostosowując się do typowych dla wizji wszechogarniającej artefaktów spowodowanych przerzutowaniem informacji między widokami. Zmiany względem oryginalnej metryki SSIM zaznaczono kolorem zielonym:

$$\boldsymbol{\mu}_c^J(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{J}_c(\boldsymbol{i', j'})] \,, \tag{3}$$

$$\boldsymbol{\sigma}_c^J(x,y) = \sqrt{\sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\boldsymbol{\omega}(i-x, j-y) \cdot \left(\boldsymbol{J}_c(\boldsymbol{i', j'})\right)^2\right] - \left(\boldsymbol{\mu}_c^J(x,y)\right)^2} \,, \tag{4}$$

$$\boldsymbol{\sigma}_c^{\boldsymbol{I\to J}}(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} [\boldsymbol{\omega}(i-x, j-y) \cdot \boldsymbol{I}_c(i,j) \cdot \boldsymbol{J}_c(\boldsymbol{i', j'})] + - \boldsymbol{\mu}_c^I(x,y) \cdot \boldsymbol{\mu}_c^J(x,y) \,, \tag{5}$$

gdzie:

$$i' = i + s_x(i,j) \,, \quad j' = j + s_y(i,j) \,, \tag{6}$$

a $s_x(i,j)$ i $s_y(i,j)$ określają względną pozycję takiego piksela w obrazie J, który znajduje się w otoczeniu o rozmiarze $B \times B$ wokół piksela (i,j), który jest najbardziej podobny do piksela (i,j) w obrazie I. Domyślnie $B = 2$.

Poszukiwanie najbardziej podobnego piksela w otoczeniu $B \times B$ skutkuje brakiem symetrii wyznaczania kowariancji, przez co konieczne jest zaznaczenie, w którym kierunku kowariancja ta jest wyznaczana. Stąd też oznaczenie $\boldsymbol{\sigma}^{I\to J}$ zamiast typowego $\boldsymbol{\sigma}^{I,J}$ używanego w przypadku SSIM.

W drugim kroku porównywane są trzy podstawowe właściwości obrazów I i J: luminancja $\boldsymbol{L}$, kontrast $\boldsymbol{C}$ i struktura $\boldsymbol{S}$:

$$\boldsymbol{L}_c^{\boldsymbol{I\to J}}(x,y) = \frac{2 \cdot \mu_c^I(x,y) \cdot \left[\mu_c^J(x,y) + s_c^{I\to J}\right] + C_1}{\mu_c^I(x,y)^2 + \left[\mu_c^J(x,y) + s_c^{I\to J}\right]^2 + C_1} \,, \tag{7}$$

$$\boldsymbol{C}_c^{\boldsymbol{I\to J}}(x,y) = \frac{2 \cdot \sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_2}{\sigma_c^I(x,y)^2 + \sigma_c^J(x,y)^2 + C_2} \,, \tag{8}$$

$$\boldsymbol{S}_c^{\boldsymbol{I\to J}}(x,y) = \frac{\sigma_c^{I\to J}(x,y) + C_3}{\sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_3} \,, \tag{9}$$

gdzie C_1, C_2 i C_3 są stałymi zapewniającymi numeryczną stabilność, zdefiniowanymi jako:

$$C_1 = (K_1 \cdot (2^b - 1))^2, \quad C_2 = (K_2 \cdot (2^b - 1))^2, \quad C_3 = \frac{C_2}{2} \,, \tag{10}$$

$$K_1 = 0.01, \quad K_2 = 0.03, \tag{11}$$

a $s_c^{I\to J}$ jest globalną różnicą składowej barwnej c obrazów I i J, wyznaczaną jako:

[1] Warto zwrócić uwagę na niefortunny dobór nazwy, który w kontekście wizji może być mylący – luminancja zdefiniowana przez autorów metryki nie ma nic wspólnego z luminancją (lumą) rozumianą jako składowa barwna.

$$s_c^{I\to J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c-1} \sum_{x=0}^{W_c-1} \left(\boldsymbol{I}_c(x,y) - \boldsymbol{J}_c(x,y) \right), \quad (12)$$

gdzie W_c i H_c są – odpowiednio – szerokością i wysokością obrazu (a ściślej – jego składowej barwnej c).

Wyznaczone właściwości $\boldsymbol{L}$, $\boldsymbol{C}$ i $\boldsymbol{S}$ są w następnym kroku wymnażane w celu wyznaczenia lokalnego podobieństwa pomiędzy obrazami I i J w punkcie (x, y):

$$\boldsymbol{Q}_c^{I\to J}(x,y) = \boldsymbol{L}_c^{I\to J}(x,y) \cdot \boldsymbol{C}_c^{I\to J}(x,y) \cdot \boldsymbol{S}_c^{I\to J}(x,y). \quad (13)$$

Następnie, podobieństwo to jest uśredniane po całej powierzchni obrazu:

$$SSIV_c^{I\to J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c-1} \sum_{x=0}^{W_c-1} \boldsymbol{Q}_c^{I\to J}(x,y) \quad (14)$$

i wszystkich składowych barwnych (luminancji i dwóch chrominancji):

$$SSIV_{YUV}^{I\to J} = \frac{SSIV_Y^{I\to J} \cdot w_Y + SSIV_U^{I\to J} \cdot w_U + SSIV_V^{I\to J} \cdot w_V}{w_Y + w_U + w_V}, \quad (15)$$

gdzie wartość wag w_Y, w_U i w_V ustalona jest tak, jak w przypadku miary IV-PSNR [3] i wynosi – odpowiednio – $w_Y = 4$, $w_U = 1$, $w_V = 1$.

Ostateczna miara podobieństwa strukturalnego pary obrazów I i J wyznaczana jest jako:

$$SSIV_{YUV}^{I,J} = \min\left(SSIV_{YUV}^{I\to J}, SSIV_{YUV}^{J\to I}\right), \quad (16)$$

co zapewnia symetryczność zaproponowanej metryki, znacząco zwiększając jej użyteczność i łatwość użycia.

4. EKSPERYMENT

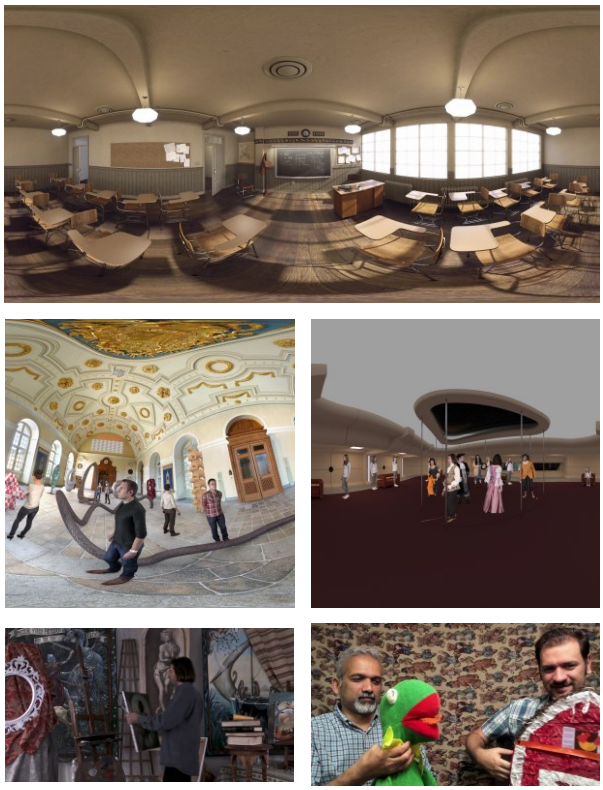
Efektywność zaproponowanej metryki została oceniona na bazie wyników konkursu grupy ISO/IEC MPEG na nowy kodek wizji wszechogarniającej [6], które to wyniki obejmowały subiektywną i obiektywną ocenę pięciu zróżnicowanych sekwencji wielowidokowych (rys. 3) zakodowanych z użyciem siedmiu różnych technik kodowania wizji wszechogarniającej. Szczegółowa metodologia eksperymentu dostępna jest w rozdziale VII artykułu opisującego metrykę IV-PSNR [3].

Zaproponowana metryka została porównana z IV-PSNR, a także SSIM i PSNR. Testowane metryki porównano wykorzystując dwa popularne współczynniki korelacji bazujące na kolejności (*rank-order coefficients*): SROCC i KROCC [7].

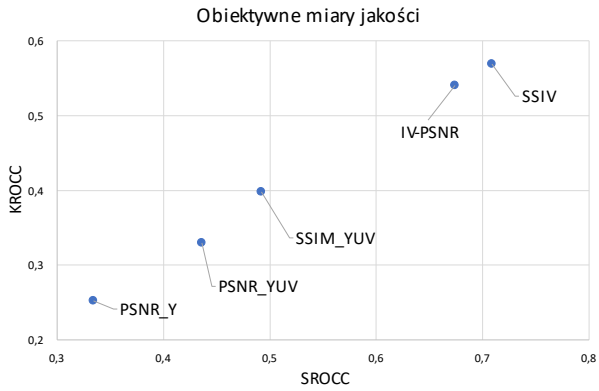
Jak pokazano na rysunku 4, zaproponowana metryka SSIV wykazuje większą korelację z wynikami testów subiektywnych, aniżeli pozostałe testowane metryki, w tym IV-PSNR. Biorąc pod uwagę fakt, iż kompleksowe badania opisane w [3] wykazały, iż inne opisane w literaturze obiektywne metryki jakości radzą sobie z poprawną oceną jakości wizji wszechogarniającej zdecydowanie gorzej niż IV-PSNR, stwierdzić można, iż SSIV przewyższa je wszystkie.

Poza porównaniem z innymi metrykami, eksperymentalnie sprawdzono również efektywność czterech wariantów estymacji metryki SSIV (oraz SSIM).

Miara SSIM opisana w artykule [15] nie jest do końca jednoznacznie zdefiniowana. Dodatkowo, sama implementacja SSIM, pozostawia wiele pytań o czym świadczy obecność publikacji wyjaśniających istotne detale i pominięte szczegóły, takich jak [8] czy [14].



Rys. 3. Zbiór sekwencji testowych.



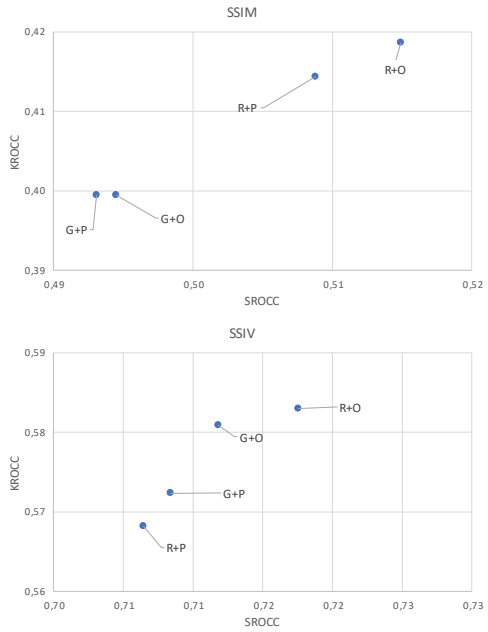
Rys. 4. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie dla pięciu testowanych metryk.

Jednym z istotnych kwestii związanych z obliczaniem miary SSIM (i jej pochodnych) jest problem skrajnych próbek obrazu. Zastosowanie maski $\boldsymbol{\omega}$ o rozmiarze 11x11 powoduje konieczność rozszerzenia obrazu o dodatkowe marginesy (ang. *padding*) co potencjalnie może mieć niekorzystny wpływ na skuteczność metryki.

Drugim problemem w wyznaczaniu SSIM jest złożoność obliczeniowa – zastosowanie maski $\boldsymbol{\omega}$ o rozmiarze 11x11 powoduje konieczność wykonania 242 operacji mnożenia na każdy punkt (ok 1,5 miliarda operacji mnożenia na obraz FullHD – 1920x1080).

W związku z powyższym zdecydowaliśmy się przebadać efekty zastosowania dwóch uproszczeń: 1. pomijania punktów dla których maska $\boldsymbol{\omega}$ sięga poza oryginalny obszar obrazu (wariant O) vs użycie rozszerzonego obrazu (wariant P); 2. zastąpienie maski Gaussowskiej (wariant G) maską prostokątną (wariant R).

Jak pokazano na rysunku 5 zastosowanie opisanych uproszczeń (a w szczególności wariantu R+O) nie tylko nie obniżyło sprawności miar SSIM i SSIV ale wręcz minimalnie poprawiło korelację z MOS przy znaczącej kilkunastoprocentowej redukcji czasu obliczeń.



Rys. 5. Korelacja pomiędzy subiektywną oceną jakości a jakością mierzoną obiektywnie dla czterech testowanych wariantów metryk SSIM i SSIV.

5. PODSUMOWANIE

W artykule przedstawiono nową metrykę oceny jakości sekwencji wizji wszechogarniającej. Nowa metryka powstała w sposób analogiczny do tego jak opracowano miarę IV-PSNR na podstawie miary PSNR. Tym razem zamiast PSNR wykorzystano SSIM.

Przeprowadzone eksperymenty pokazały, że zaproponowana miara SSIV jest jeszcze lepiej skorelowana z wynikami oceny subiektywnej (MOS) niż inne znane metryki, w tym IV-PSNR. Co więcej, w pracy przedstawiono wyniki dla czterech wariantów wyznaczania miary SSIM, a w konsekwencji także SSIV, dążące do zredukowania złożoności obliczeniowej przedstawionej propozycji.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego.

LITERATURA

[1] Boyce Jill, Dore Renaud, Dziembowski Adrian, Fleureau Julien, Jung Joel, Kroon Bart, Salahieh Basel, Kumar Malamal Vadakital Vinod, Yu Lu. 2021. "MPEG Immersive Video coding standard". *Proceedings of the IEEE*, 119 (9): 1521-1536.

[2] Domański Marek, Stankiewicz Olgierd, Wegner Krzysztof, Grajek Tomasz. 2017. "Immersive visual media — MPEG-I: 360 video, virtual navigation and beyond". *24th International Conference on Systems, Signals and Image Processing*, Poznań, Polska.

[3] Dziembowski Adrian, Mieloch Dawid, Stankowski Jakub, Grzelka Adam. 2022. "IV-PSNR – The objective quality metric for immersive video applications". *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11): 7575-7591.

[4] Dziembowski Adrian, Mieloch Dawid. 2022. "Nowe techniki kompresji wizji dla rzeczywistości wirtualnej – MPEG Immersive Video". *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne.* 4: 116-123.

[5] Han Jingning i in. 2021. "A technical overview of AV1". *Proceedings of the IEEE* 109 (9): 1435-1462.

[6] ISO/IEC. 2019. "Call for Proposals on 3DoF+ Visual". *ISO/IEC JTC1/SC29/WG11 MPEG N18145*, Marakesz, Maroko.

[7] Myers Jerome, Well Arnold. 2003. "Research design and statistical analysis". *Lawrence Erlbaum Associates*, Londyn, UK.

[8] Nilsson Jim, Akenine-Möller Tomas. 2020. "Understanding SSIM". arxiv.org/abs/2006.13846.

[9] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2015. „DIBR synthesized image quality assessment based on morphological pyramids". *3DTV-CON*, Lizbona, Portugalia.

[10] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2015. "DIBR synthesized image quality assessment based on morphological wavelets". *International Workshop on Quality of Multimedia Experience QoMEX*, Costa Navarino, Grecja.

[11] Sandić-Stanković Dragana, Kukolj Dragan, Le Callet Patrick. 2016. "Multi–Scale Synthesized View Assessment Based on Morphological Pyramids". *Journal of Electrical Engineering* 67 (1): 1–9.

[12] Sullivan Gary J., Ohm Jens-Rainer, Han Woo-Jin, Wiegand Thomas. 2012. "Overview of the High Efficiency Video Coding (HEVC) Standard". *IEEE Transactions on Circuits and Systems for Video Technology* 22: 1649–1668

[13] Sullivan Gary J., Wiegand Thomas. 2005. "Video Compression – From Concepts to the H.264/AVC Standard". *Proceedings of the IEEE* 93 (1): 18-31.

[14] Venkataramanan Abhinau K., Wu Chengyang, Bovik Alan C., Katsavounidis Ioannis, Shahid Zafar. 2021. "A Hitchhiker's Guide to Structural Similarity". IEEE Access 9: 28872-28896.

[15] Wang Zhou, Bovik Alan C., Sheikh Hamid, Simoncelli Eero. 2004. "Image quality assessment: From error measurement to structural similarity". *IEEE Transactions on Image Processing*, 13 (4): 600-612.

[16] „Koder AOM", dostępny [02.06.2024]: https://aomedia.googlesource.com/aom/

[17] „Koder x264", dostępny [02.06.2024]: https://www.videolan.org/developers/x264.htm

[18] „Koder x265", dostępny [02.06.2024]: https://x265.com/