Olgierd Stankiewicz
Krzysztof Wegner
Tomasz Grajek
Marek Domański

Katedra Telekomunikacji Multimedialnej i Mikroelektroniki
Politechniki Poznańskiej
ostank@multimedia.edu.pl

DOI: 10.15199/59.2018.6.55

KKRRiT 2018
Gdańsk, 20-22 czerwca 2018

NOWE MEDIA IMMERSYJNE

NEW IMMERSIVE MEDIA

Streszczenie: **W niniejszym artykule rozważamy immersyjne (wszechogarniające) media wizyjne, które są obecnie badane w środowisku naukowym. Obejmują one dobrze ugruntowane technologie, takie jak obraz dookólny, a także te, które są intensywnie rozwijane, jak systemy swobodnej nawigacji. Z pomocą przykładów, opisujemy również cechy immersyjnych mediów wizyjnych, które odróżniają je od klasycznego ruchomego obrazu 2D. Ponadto krótko przedstawiamy technologie kompresji, które są obecnie rozpatrywane w kontekście projektu standaryzacyjnego MPEG-I rozpoczętego przez ISO/IEC.**

Abstract: **In this paper we present a survey of immersive visual media that are currently considered within research community. These include the well-established technologies, like 360-degree panoramas, as well as those being intensively developed like free viewpoint video. On the examples of such systems, we show the features of the immersive visual media that distinguish them from the classical 2D video. Also, we present the compression technologies that are currently considered in the context of standardization of the immersive visual media in the MPEG-I project recently launched by ISO/IEC.**

Słowa kluczowe: **Telewizja swobodnego punktu widzenia, MPEG-I, obraz dookólny, obraz wszechkierunkowy, media immersyjne.**

Keywords: **Immersive visual media; Free viewpoint video; 3D 360 video; MPEG-I.**

1. WSTĘP

Nazwa nowego rodzaju mediów, zwanych immersyjnymi (lub wszechogarniającymi), pochodzi od łacińskiego czasownika immergere, co oznacza zanurzenie się lub zanurzenie w czymś. W przypadku mediów cyfrowych jest to termin określający zdolność systemu technicznego do całkowitego „wchłonięcia" użytkownika w przedstawioną rzeczywistość. Multimedia immersyjne [1] mogą być związane zarówno z treścią naturalną, jak i generowaną komputerowo. W pracy skoncentrujemy się na treści naturalnej, która zastała zarejestrowana za pomocą kamer, mikrofonów i ewentualnie jest uzupełniana danymi z czujników dodatkowych, takich jak kamery głębi. Immersja ma różne aspekty, związane ze sposobami przekonywania naszych ludzkich zmysłów, że jesteśmy obecni w przedstawionej treści. W tym artykule koncentrujemy się na tych aspektach, związanych tylko z wizją.

Przykładami mediów immersyjnych są rzeczywistość wirtualna, rzeczywistość mieszana i rzeczywistości rozszerzona, które ostatnimi czasy są bardzo dynamicznie rozwijane. Obecnie najbardziej obiecującymi urządzeniami do prezentacji obrazu, który pozwala na całkowite wchłonięcie widza są gogle VR (ang. Head Mounted Display), umieszczane na głowie i wyświetlające obraz bezpośrednio na wyświetlaczach umieszczonych wprost przed oczami widza. W ostatnich latach obserwuje się ciągły rozwój tego typu urządzeń w różnych skalach: od prostych opartych na smartfonach (Google Cardboard [2]) po dedykowane urządzenia współpracujące z komputerami PC (np. Oculus Rift) [3]. Rozwój i popularyzacja urządzeń umożliwiających immersję przyśpieszyły rozwój rynku i zintensyfikowały prowadzone badania. Obecnie trwają prace nad wieloma technologiami związanymi z immersją, z których każdy prezentuje różne poziomy/zakresy zanurzenia widza w prezentowany świat (realizm). Niektóre są bardzo ograniczone, np. pozwalają użytkownikowi zmienić tylko kierunek widzenia [4,5], podczas gdy inne są bardzo zaawansowane, np. pozwalają użytkownikowi swobodnie przemieszczać się w prezentowanym świecie [6,7] lub wręcz wchodzić w interakcje z prezentowanym środowiskiem [8, 9].

Dobrym przykładem takiej interaktywnej treści jest obraz przestrzenny wraz z towarzyszącym mu dźwiękiem przestrzennym, który pozwala człowiekowi wirtualnie eksplorować nowe, nieznane miejsca, nie zawsze przyjaznych dla odwiedzających. Podczas wirtualnego spaceru piechur nie naraża się na niebezpieczeństwo i może wybrać wirtualną trajektorię spaceru, może przystanąć i rozejrzeć się, usłyszeć dźwięki dżungli, wiatru itp. W takich zastosowaniach, odpowiednią treść uzyskuje się przez użycie skupisk kamer i mikrofonów, a po nagraniu zebrany materiał musi zostać przetworzony w celu odpowiedniej całościowej reprezentacji sceny audiowizualnej. Prezentacja takich treści wymaga głównie mechanizmów renderowania, np. w celu produkcji obrazu i dźwięku, które odpowiadają określonej lokalizacji i kierunkowi obserwacji aktualnie wybranym przez wirtualnego eksploratora dżungli. Dlatego prezentowanie takiej treści można również sklasyfikować jako prezentację rzeczywistości wirtualnej, mimo że cała zawartość prezentuje rzeczywiste obiekty w ich rzeczywistych lokalizacjach i ruchach. Niemniej jednak treści generowane komputerowo (zarówno samodzielne, jak i mie-

szane z naturalnymi treściami) będą wyłączone z zakresu niniejszego artykułu.

Różnorodność pojawiających się technologii zapewniających immersję zmotywowała komitety normalizacyjne do podjęcia niezbędnych kroków. W szczególności komitet MPEG (ang. Motion Picture Experts Group) przy Międzynarodowej Organizacji Normalizacyjnej (ISO) oraz Międzynarodowej Komisji Elektrotechnicznej (IEC) uruchomił w zeszłym roku nowy projekt [10,11] nazwany MPEG-I. Przyszła norma MPEG-I, zatytułowana "Coded Representation of Immersive Media", ma na celu normalizację nie tylko obecnie znanych rozwiązań, np. obraz dookólny (ang. 360 degree video), który umożliwia percepcje obrazu z 3 stopniami swobody (kąty RPY), ale także przyszłe system, np. systemy o sześciu stopniach swobody (6 DoF), w który widz może nie tylko się rozglądać, ale i swobodnie przemieszczać.

W niniejszym artykule przedstawiamy aktualny stan techniki w dziedzinie immersyjnych mediów wizualnych, w tym obrazu 360, stereoskopowego obrazu 360, oraz dwuocznego trójwymiarowego obrazu 360 [12], systemów swobodnej nawigacji i systemów opartych o chmury punktów, zarówno w kontekście prac badawczych, jak i normalizacji.

2. CECHY WIZJI IMMERSYJNEJ

Aby wchłonąć użytkownika całkowicie w prezentowany świat, zastosowana technologia musi przekonać nasze zmysły, iż ten świat jest realny. Wyświetlenie widzowi obrazu przed oczami to za mało, aby w wystarczającym stopniu oszukać ludzki mózg. Stopień immersji można jednak zwiększyć, jeśli uwzględnione zostaną następujące cechy ludzkich zmysłów:

Swoboda rozglądania się. Możliwość swobodnego rozglądania się we wszystkich kierunkach z trzema stopniami swobody (3 DoF), umożliwia ludzkiemu mózgowi zbudowanie holistycznego modelu otaczającego świata. Ten proces ma kluczowe znaczenie, aby zapewnić w pełni immersyjne doświadczenie.

Swoboda przemieszczania. Zdolność użytkownika do swobodnego poruszania się w prezentowanym świecie we wszystkich kierunkach (3 DoF) znacząco zwiększa poziom realizmu i zaangażowanie użytkownika, poprzez obserwację paralaksy ruchu i umożliwienie eksplorowania prezentowanego świata.

Opóźnienie. Ludzki mózg jest bardzo wrażliwy na różnice w czasie postrzegania informacji pochodzących z różnych zmysłów. Niedopasowanie momentu percepcji obrazu i momentu przemieszczenia powoduje symptomy kinetozy, choroby zbliżonej do choroby lokomocyjnej, których można uniknąć minimalizując opóźnienia systemu.

Widzenie obuoczne. Ludzki system wzrokowy wykorzystuje informacje z obu oczu, aby dostrzec głębię sceny. Bez właściwych różnic w obrazach wyświetlanych lewemu i prawemu oku scena jest postrzegana jako płaska i nienaturalna.

Rozdzielczość. Ekrany zamontowane w hełmach VR są bardzo blisko oczu użytkownika i dlatego liczba punktów musi być bardzo wysoka, aby uniknąć aliasingu. Jest to szczególnie ważne, gdy użytkownik porusza się lub nieznacznie obraca głową, co w przypadku zbyt niskiej rozdzielczości wyświetlacza powoduje nienaturalne przeskoki krawędzi postrzeganego obrazu.

Uosobienie, personalizacja. Bardzo ważna jest właściwa percepcja własnego ja w prezentowanym środowisku. W szczególności zdolność widzenia części własnego ciała przekonuje użytkownika o byciu częścią prezentowanej rzeczywistości.

Interaktywność. Umożliwienie użytkownikowi manipulowania obiektami zapewnia silne przesłanki dotyczące integralności prezentowanej rzeczywistości.

3. TECHNOLOGIE IMMERSYJNE

W rozdziale tym zostaną zaprezentowane istniejące i powstające technologie, które zapewniają immersję użytkownika. Nie wszystkie z nich obsługują wszystkie wymienione wcześniej cechy, co oczywiście obniża osiągnięty poziom realizmu.

3.1. Wizja dookólna

Wizja dookólna przekazuje widok sceny we wszystkich kierunkach widziany z danego punktu (Rys. 1). W praktyce obraz dookólny rejestrowany jest najczęściej za pomocą specjalnych kamer dookólnych złożonych z zestawu od 4 do 6 kamer patrzących w rozbieżnych kierunkach (Rys. 2). Obrazy z poszczególnych kamer są następnie łączone razem [13] w celu utworzenia pojedynczego widoku panoramy.

Rys. 1. Przykładowy obraz dookólny

Rys. 2. Przykładowa kamera dookólna zbudowana z 6 kamer szerokokątnych

Dane obrazu dookólnego są reprezentowane w formie przypominającej klasyczny obraz dwuwymiarowy, z tą różnicą, że współrzędne punktu obrazu reprezentują kierunek obserwacji (kąt), a nie położenie na płaskiej płaszczyźnie obrazu kamery. Odległość między lewą i prawą krawędzią obrazu wynosi 360 stopni, a zatem punkty położone na przeciwległych krawędziach, w rzeczywistości, sąsiadują ze sobą na powierzchni sfery. Odwzorowanie współrzędnych długości i szerokości geograficznej na powierzchni sfery na płaszczyznę obrazu dwuwymiarowego może być określone na różne sposoby. Najpowszechniej wykorzystywane są prze-

kształcenie walcowe równoodległościowe i centralne przekształcenie walcowe [14], oba mające zalety i wady związane z przedstawionym zakresem i rozdzielczością kątów. W modelu, akwizycja każdej kolumny punktów dokonywana jest przez osobną kamerę z bardzo wąskim poziomym polem widzenia (FoV) i pewnym pionowym polem widzenia. W przypadku rzutu prostopadłościennego pionowa wartość FoV wynosi 180 stopni, zaś w przypadku rzutowania perspektywicznego cylindrycznego jest mniejsza.

Różnorodność możliwych odwzorowań powierzchni sfery na płaszczyznę jest wyzwaniem dla normalizacji [14]. Jednym z obecnie rozważanych rozwiązań jest wykorzystanie odwzorowania opartego na siatce punktów węzłowych zamiast zestawu wybranych formuł matematycznych. Prace w ramach grupy MPEG są nadal w toku, ale norma dla obrazu dookólnego znana pod angielską nazwą Omnidirectional Media Application Format - OMAF ma zostać ukończona w pierwszej połowie 2018 roku.

Prezentując wybrany fragment obrazu dookólnego, można pozwolić użytkownikowi na swobodny wybór kierunku obserwacji sceny (obrót). Użytkownik nie może jednak ruszyć się z miejsca. Dlatego też obraz dookólny określany jest często, jako obraz z trzeba stopniami swobody (ang. 3 Degrees of Freedom – 3 DoF). Ponadto, ponieważ oboje oczu obserwuje ten sam wycinek panoramy, nie ma wrażenia głębi obrazu.

3.2. Stereoskopowy obraz dookólny

Stereoskopowe obrazy dookólne to rozszerzenie idei obrazu dookólnego, w którym rejestrowane są dwie panoramy sceny - jedna dla lewego, a druga dla prawego oka. Często spotykana nazywa to "trójwymiarowy obraz dookólny". Obie panoramy zazwyczaj złożone są w jeden obraz, w którym u góry znajduje się obraz dla lewego oka a u dołu dla prawego (rys. 3).

Rys. 3. Para panoram z sekwencji „Dancer 360" [15] tworząca stereoskopowy obraz dookólny

Wykorzystanie dwóch panoram umożliwia prezentację różnych obrazów dla lewego i prawego oka, co wywołuje u widza wrażenie głębi sceny. Niestety doznania głębi są ograniczone do obszarów w pobliżu równika rzutu, ponieważ na biegunach oba obrazy są takie same. Ponadto użytkownik nie może się poruszać, nawet nieznacznie, co, przy poruszaniu głową, powoduje znaczący dysonans poznawczy pomiędzy tym, co użytkownik widzi, a tym, co czuje ze zmysłu równowagi.

Oczekuje się, że normy dotyczące stereoskopowych obrazów dookólnych powstaną do końca 2018 roku.

3.3. Dwuoczny, trójwymiarowy obraz dookólny

Celem dwuocznej trójwymiarowej technologii obrazu dookólnego (ang. binocular 3D 360 video) jest rozwiązanie największych ograniczeń stereoskopowego obrazu dookólnego: ograniczonego odczuwania głębi poza obszarem równika oraz braku paralaksy ruchu przy ruchach głowy użytkownika. Zakłada się, że użytkownik nie będzie mógł przemieszczać się swobodnie, a może jedynie nieznacznie poruszać głową, to ta technologia często określana jest jako 3 DoF+ (plus) - nie zapewnia pełnego zestawu sześciu stopni swobody (obrotów i swobodnego przemieszczania się).

Renderowanie widoków z paralaksą ruchu wymaga pewnych informacji o głębi sceny, np. dalsze obiekty poruszają się pozornie wolniej przy zmianie perspektywy obserwacji niż obiekty bliższe. Obecnie w ramach grupy MPEG rozważane są dwa rozwiązania techniczne umożliwiające realizacji systemu 3DoF+. Pierwszym z nich jest zastosowanie warstwowego stereoskopowego obrazu dookólnego, w którym każdej warstwie przypisywany jest stały poziom głębi. Przy ruchach głowy widza, użytkownik obserwuje scenę z pozycji nie centralnej, a więc pozornie kolejne warstwy obrazu przesuwają się poprawiając wrażenie głębi.

Drugim rozważanym rozwiązanie jest wykorzystanie map głębi, które przekazują informacje o odległości poszczególnych punktów sceny od powierzchni sfery. Oczywiście wymaga to informacji o głębi sceny, która musi zostać uzyskana bezpośrednio (np. za pomocą kamer głębi) lub zostać wyliczona algorytmicznie za pomocą technik estymacji rozbieżności.

Istnieją już prace, które opisują techniki pozwalające na wyznaczenie głębi ze stereoskopowego obrazu dookólnego. Np. w pracy [16] pokazano, że możliwe jest wyznaczenie głębi stereoskopowego obrazu dookólnego za pomocą klasycznych algorytmów estymacji głębi. Oczekuje się, że trójwymiarowe obrazu dookólnego zostaną znormalizowane do końca 2019 roku.

3.4. Telewizja swobodnego punktu widzenia

W systemie swobodnego punktu widzenia użytkownik może dowolnie wybierać punkt obserwacji sceny, to znaczy zarówno kierunek patrzenia i swoje położenie w przestrzeni. Dlatego takie systemy zapewniają użytkownikowi sześć stopni swobody (6 DoF). Różnica między trójwymiarowym obrazem dookólnym i systemami swobodnej nawigacji jest w formacie zapisu obrazu. Systemy telewizji swobodnego punktu widzenia wykorzystują format reprezentacji w postaci obrazów wielowidokowych wraz z mapami głębi (ang. Multiview Video plus Depth - MVD). W MVD scena jest rejestrowana za pomocą ograniczonej liczby kamer (np. 10) umieszczonych wokół sceny (Rys. 4). Każdy widok zarejestrowany przez kamerę jest powiązany z odpowiednią mapą głębi.

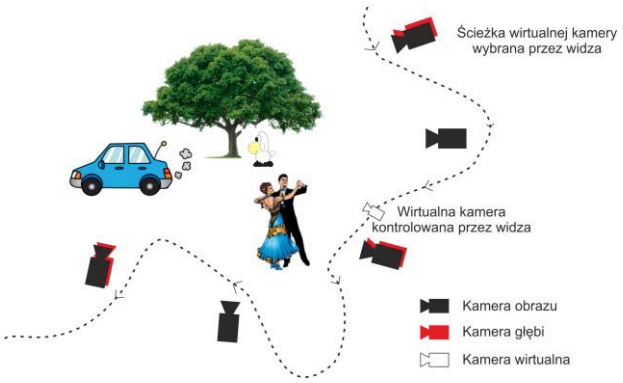
Widoki i mapy głębi łącznie używane są do renderowania (syntezowania) pożądanego widoku dla lewego i prawego oka użytkownika, np. za pomocą techniki renderowania widoków na podstawie głębi i obrazu (ang. Depth-Image Based Rendering - DIBR).

Oczywiście, poruszanie się w systemie FTV ograniczone jest do obszaru, który został zarejestrowany przez kamery. Uzyskiwane u widza wrażanie przypomina oglądanie sceny przez czyste okno. Z tego względu systemy swobodnego punktu widzenia w żargonie używanym w grupie MPEG nazywane są "okienkowymi

6DoF" (Windowed 6 Degrees of Freedom) (Rys. 5). Ogranicza to również swobodę użytkownika w praktycznych zastosowaniach. Na przykład, jeśli kamery patrzą na zewnątrz, wówczas, podobnie jak w przypadku dwuocznego, trójwymiarowego obrazu dookólnego, dozwolony jest tylko bardzo mały ruch użytkownika. Z drugiej strony, jeśli kamery znajdują się tylko po jednej stronie sceny, użytkownika może poruszać się niemal bez ograniczeń, ale możliwość obrotu jest bardzo ograniczona.

Ze względu na wyzwania w tej dziedzinie, norm dotyczących systemów swobodnego widzenia zapewne nie uda opracować się przed rokiem 2021.

Rys. 4. Eksperymentalny system telewizji swobodnego punktu widzenia wybudowany na Politechnice Poznańskiej



Rys. 5. Ilustracja działania systemu swobodnej nawigacji

4. PODSUMOWANIE

W artykule opisaliśmy różne wizyjne media immersyjne, w tym obrazy 360, stereoskopowe obrazy 360, dwuoczne trójwymiarowe obrazy 360 i systemy swobodnej nawigacji. Dla każdego z prezentowanych przykładów rozważaliśmy obsługiwane funkcje, które określają osiągnięty poziom zanurzenia. Jak pokazano, poziom ten różni się wśród rozważanych technologii i zmienia się w zależności od ich złożoności. Oczekuje się, że niektóre z nich będą dostępne w niedalekiej przyszłości. Fakt ten jest jedną z motywacji prac grupy ISO/IEC MPEG w projekcie MPEG-I, którego celem jest standaryzacja immersyjnych mediów wizyjnych. Projekt ten ma być realizowany w fazach: technologie, które są już dostępne, będą standaryzowane najpierw, a następnie udostępniane będą rozszerzenia związane z technologiami, które opracowane zostaną później.

Praca finansowana ze środków na naukę, jako projekt DS „Działalność Statutowa”.

LITERATURA

[1] F. Isgro, E. Trucco, P. Kauff, O. Schreer, “*Three-dimensional image processing in the future of immersive media”, IEEE Trans Circuits Syst. Video Techn.,* vol. 14, 2004, pp. 288 – 303.
[2] https://vr.google.com/cardboard/ - avaiable April 2017.
[3] https://www.oculus.com/rift/ - avaiable April 2017.
[4] K. C. Huang, P. Y. Chien, et Al., *"A 360-degree panoramic video system design" Technical Papers of 2014 International Symposium on VLSI Design, Automation and Test,* Hsinchu, 2014, pp. 1-4.
[5] T. M. Liu, C. C. Ju, at Al. *"A 360-degree 4K×2K panoramic video recording over smart-phones," 2016 IEEE International Conference on Multimedia & Expo Workshops* (ICMEW), Seattle, WA, 2016, pp. 1-1.
[6] M. Tanimoto, M. P. Tehrani, T. Fujii, T. Yendo *“FTV for 3-D spatial communication”, Proc. IEEE*, vol. 100, no. 4, pp. 905-917, 2012.
[7] M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, *“New results in free-viewpoint television systems for horizontal virtual navigation”, 2016 IEEE International Conference on Multimedia and Expo* ICME 2016, Seattle, USA, 2016.
[8] A. J. Fairchild, S. P. Campion, A. S. García, R. Wolff, T. Fernando and D. J. Roberts, *"A Mixed Reality Telepresence System for Collaborative Space Operation," in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 814-827, April 2017.
[9] C. Schwede and T. Hermann, *"HoloR: Interactive mixed-reality rooms," 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom),* Gyor, 2015, pp. 517-522.
[10] *“New Work Item Proposal on Coded Representation of Immersive Media”, ISO/IEC JTC1/SC29/WG11* MPEG2016/N16541 Chengdu, CN – October 2016.
[11] G. Lafruit, M. Domański, K. Wegner, et Al., *“New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV”, IST Electronic Imaging, Stereoscopic Displays and Applications* XXVII, San Francisco 2016.
[12] S. Li. *“Binocular spherical stereo” IEEE Trans. on Intelligent Transportation Systems,* 9(4):589–600, 2008.
[13] M. Z. Bonny, M. S. Uddin, *"Feature-based image stitching algorithms," 2016 International Workshop on Computational Intelligence* (IWCI), Dhaka, 2016, pp. 198-203.
[14] Yan Ye, Elena Alshina, Jill Boyce, *“Algorithm descriptions of projection format conversion and video quality metrics in 360Lib” Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11* 6th Meeting: Document: JVET-F1003-v1, Hobart, AU, 31 March – 7 April 2017.
[15] G. Bang, G. S. Lee, N. H. Hur, *“Test materials for 360 3D video application discussion”, ISO/IECJTC1/SC29/WG11* MPEG2016/M37810 February 2016, San Diego, USA
[16] K. Wegner, O. Stankiewicz, T. Grajek, M. Domański *“Depth estimation from circular projection of 360 degree 3D video” ISO/IEC JTC1/SC29/WG11* MPEG2017/m40596, April 2017, Hobart, Australia.