

WYKORZYSTANIE WEJŚCIOWYCH MAP GŁĘBI W ESTYMACJI GŁĘBI PO STRONIE DEKODERA THE USE OF INPUT DEPTH MAPS IN DECODER-SIDE DEPTH ESTIMATION

Dominika Klóska, Błażej Szydełko, Dawid Mieloch
Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
dominikkloska@gmail.com, szydelkob@hotmail.com, dawid.mieloch@put.poznan.pl

DOI: 10.15199/59.2022.4.65

Streszczenie: Artykuł opisuje usprawnienia procesu estymacji map głębi po stronie dekodera wizji wszechogarniającej. W prezentowanej propozycji mapy głębi są przesyłane tylko dla części widoków, umożliwiając znaczące zmniejszenie czasu wymaganego na estymację geometrii przy jednoczesnym zwiększeniu jej jakości. Rozwiązanie obejmuje modyfikację oprogramowania ISO/IEC MPEG służącego do wyznaczania geometrii sceny oraz implementacji kodera MPEG Immersive Video. Wyniki testów obejmują porównanie czasu dekodowania i uzyskiwanych prędkości bitowych.

Abstract: This paper presents a proposal to improve the depth estimation process performed in decoder-side depth estimation. In the proposed solution, depth maps are sent only for a part of the views, allowing a significant reduction in the time required for geometry estimation and a simultaneous increase of its quality. The solution includes the modification of the ISO/IEC MPEG Video Coding software for depth estimation and the implementation of the MPEG Immersive Video encoder. The test results consist of the comparison of decoding times and required bitrates.

Słowa kluczowe: estymacja po stronie dekodera, wizja wszechogarniająca, wejściowe mapy głębi

Keywords: decoder-side depth estimation, immersive vision, input depth maps.

1. WSTĘP

Obecnie istnieje wiele systemów multimedialnych, które dają użytkownikom dostęp do trójwymiarowej wizji z dodatkową możliwością swobodnego nawigowania polem widzenia w wyświetlanej scenie [3], [9]. Wizja generowana przez te systemy ma charakteryzować się wysokim poziomem realizmu, co wiąże się z koniecznością posiadania szczegółowej informacji na temat geometrii sceny. Aby uzyskać tego typu informację należy m.in. wykorzystać wysokiej jakości widoki wejściowe w procesie estymacji map głębi [4]. Sekwencje potrzebne do stworzenia wizji wszechogarniającej są rejestrowane przez system wielokamerowy, w którym zarówno rodzaj kamer (np. dookólne) oraz ich ułożenie (np. macierz, łuk) mogą być dowolne. Istnieje zależność między liczbą kamer użytych do rejestrowania sceny trójwymiarowej oraz jakością estymowanych map głębi – im więcej kamer użyto do rejestracji sceny tym lepszą jakość ma wyznaczona głębia [8].

Jak opisano powyżej, do stworzenia wizji wszechogarniającej wymagana jest duża ilość danych, czego na-

stępstwem są wysokie prędkości bitowe wymagane do zakodowania strumienia danych. Z tego powodu, aby umożliwić przesyłanie tych danych, należy zakodować widoki i odpowiadające im mapy głębi używając metod kodowania o dużej wydajności.

Do tej pory w celu kodowania sekwencji wielowidokowych wykorzystywano kodery MV-HEVC oraz 3D-HEVC [11], które jednak nie są w stanie do końca sprostać wymaganiom, jakie stawia przed nimi nowoczesna wizja wszechogarniająca. Nie są one dostosowane do kodowania sekwencji nagrywanych systemem wielokamerowym, w którym użyte kamery były dookólne lub ich ułożenie było nieliniowe.

Przedstawione powyżej problemy oraz wzrost popularności wizji wszechogarniającej wśród społeczności specjalizującej się w przetwarzaniu wizji i obrazu sprawiły, że zaczęły pojawiać się nowe propozycje standaryzacji procesu kodowania oraz dekodowania tego typu wizji [12]. Najnowszym ze standardów kodowania stworzonym dla takich celów jest MIV – MPEG Immersive Video [1]. Jest on formą przetwarzania wstępnego, w której wybierane są takie widoki lub fragmenty widoków, które są niezbędne do wygenerowania wizji wszechogarniającej. Te dane są następnie pakowane do atlasów, czyli obrazów zawierających jeden lub więcej widoków wejściowych lub fragmenty tych widoków. Po przetwarzaniu wstępnym, MIV wykorzystuje dowolny koder wizyjny (np. HEVC [10], VVC [2]) do zakodowania stworzonych atlasów. Główny profil standardu, MIV Main, pozwala na kompresję dokonywaną poprzez usunięcie międzywidokowej informacji nadmiarowej, ale przesyłane, zakodowane mapy głębi mogą zajmować nawet ponad 50% strumienia bitów [7].

W niniejszej pracy skupiono się na profilu MIV DSDE (ang. *decoder-side depth estimation*, czyli estymacja głębi po stronie dekodera), ponieważ pozwala on znacznie zredukować prędkość bitową w porównaniu do profilu głównego, w związku z tym, że mapy głębi nie są przesyłane w strumieniu. Atlasy profilu MIV DSDE zostały przedstawione na Rys. 1, gdzie każda kolumna przedstawia osobny atlas. Jak opisano wyżej, profil ten nie przesyła głębi do dekodera, a więc nie ma atlasów zawierających geometrie. Dzięki temu zwiększona zostaje wydajność typowych koderów wizyjnych, które nie zostały stworzone do kodowania geometrii. Schemat kodera MIV został przedstawiony na Rys. 2.



Rys. 1. Atlasy wygenerowane przez profil MIV DSDE. Każda kolumna widoków odpowiada jednemu atlasowi.



Rys. 2. Schemat kodera profilu MIV Main. Na pomarańczowo zaznaczono te elementy, których nie wykorzystuje koder profilu MIV DSDE.

Estymacja głębi w procesie dekodowania w MIV nie została ujęta w standardzie [6], a więc można ją realizować dowolną metodą. Immersive Video Depth Estimation (IVDE) [8] jest oprogramowaniem referencyjnym grupy ISO/IEC MPEG do wyznaczania geometrii sceny, które zostało dostosowane do standardu MIV. Dzięki temu odpowiada wymaganiom narzucanym przez MIV na proces estymacji głębi i dlatego to właśnie to oprogramowanie zostało wykorzystane na potrzeby niniejszego artykułu. Pomimo wspomnianych zalet, wciąż istnieją pewne aspekty procesu wyznaczania geometrii sceny, które wymagają szczególnej uwagi. Jednym z nich jest estymacja głębi powierzchni płaskich znajdujących się w scenie. Gradienty tych powierzchni często są nieciągłe w wyznaczonych mapach głębi (Rys. 3), co negatywnie wpływa na subiektywną jakość wizji wszechogarniającej.



Rys. 3. Przykładowe mapy głębi z problemem braku gradientów dla powierzchni gładkich dla sekwencji Classroom (A), Kitchen (B) oraz Group (C).

Ten artykuł opisuje propozycję rozwiązania problemu estymacji głębi powierzchni gładkich w oprogramowaniu IVDE oraz wykorzystanie wprowadzonej zmiany w zmodyfikowanej implementacji kodera zrealizowanej w oprogramowaniu TMIV (Test Model for MPEG Immersive Video [1]). W prezentowanej propozycji mapy głębi są przesyłane tylko dla części widoków, umożliwiając znaczące zmniejszenie czasu wymaganego na estymację głębi przy jednoczesnym zwiększeniu jej jakości. Wyniki badań obejmują porównanie prędkości bitowych wymaganych do zakodowania sceny oraz czasu dekodowania zmodyfikowanego oprogramowania TMIV z oprogramowaniem referencyjnym. Zaproponowane

zmiany pozwoliły na znaczne skrócenie czasu dekodowania oraz zmniejszyły problem estymacji powierzchni gładkich nie zwiększając przy tym prędkości bitowych użytych na potrzeby przesyłania danych.

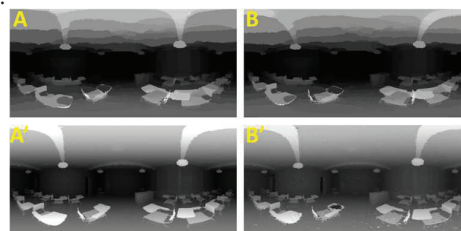
2. WEJŚCIOWE MAPY GŁĘBI

2.1. Modyfikacja oprogramowania IVDE

IVDE jest oprogramowaniem referencyjnym grupy ISO/IEC MPEG służącym do wyznaczania geometrii sceny na podstawie widoków zarejestrowanych przez system wielokamerowy. Głębina punktu może być wyznaczona jeżeli został on zarejestrowany przez dwie lub więcej kamer systemu. Dla każdego punktu z widoku wejściowego w sąsiednich widokach poszukiwany jest jego odpowiednik, czyli punkt o największym podobieństwie. Jak można zauważyć, proces poszukiwania odpowiadających sobie punktów tworzy ścisłą relację między mapami głębi sąsiadujących widoków.

Sekwencje wielowidokowe posiadają wysokiej jakości mapy głębi pozyskane wprost z programu graficznego, w którym zostały stworzone (w przypadku sekwencji wygenerowanych komputerowo), lub wyestymowane przy użyciu ręcznie dobranych parametrów. W proponowanym podejściu, zmodyfikowano oprogramowanie referencyjne tak, aby umożliwić wykorzystanie tych map głębi.

Wprowadzone zmiany zakładały, że danymi podawanymi na wejście oprogramowania IVDE będą nie tylko widoki wejściowe, ale również odpowiadające im wysokiej jakości mapy głębi. Wartości próbek mapy głębi podanej na wejście IVDE są kopiowane do odpowiadającej jej mapy wyjściowej. Ze względu na to, iż w procesie estymacji wszystkie mapy głębi są ze sobą powiązane, przesłanie jednej lub kilku wejściowych map głębi wpływa pozytywnie na wszystkie głębokości wyjściowe. Dodatkowo, proces estymacji głębi w IVDE jest oparty na małych segmentach, a nie na punktowej estymacji [8], więc umożliwia poprawę błędów wynikających z kompresji. Z tego powodu wejściowe mapy głębi mogą być silnie skompresowane.



Rys. 4. Mapy głębi dla sekwencji Classroom Video: (A) i (B) – metoda referencyjna, (A') i (B') – modyfikacja.

Na Rys. 4 pokazano jak wprowadzone zmiany wpłynęły na obecność gradientów powierzchni gładkich w wyznaczonej geometrii sceny (na wejście została podana mapa głębi odpowiadająca widokowi 0).

2.2. Modyfikacja oprogramowania TMIV

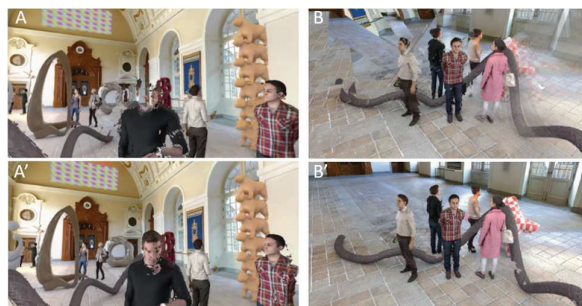
Wizja wszechogarniająca w dużej mierze składa się z widoków wirtualnych prezentujących te części sceny, które nie zostały zarejestrowane przez żadną z kamer systemu wielokamerowego. Są one syntezowane na podstawie wyznaczonej głębi, a więc istnieje ścisła zależność

między jakością wizji wszechogarniającej, a jakością wyznaczonej geometrii. Z tego powodu problem opisany w rozdziale 1 wpływa negatywnie na jakość wizji wszechogarniającej. Zważywszy na to, iż modyfikacja opisana w podrozdziale 2.1 poprawiła jakość głębi, wprowadzono kilka zmian w oprogramowaniu TMIV by móc ją wykorzystać.

Jako iż modyfikacja dotyczy profilu MIV DSDE, wykorzystywane są w niej te same widoki wejściowe, które są pakowane do atlasów w MIV DSDE. Różnica polega na kolejności, w jakiej widoki te są układane w atlasach metody zmodyfikowanej. W pierwszym atlasie w metodzie zmodyfikowanej przesyłane są widoki zawierające jak najwięcej informacji o scenie (np. w przypadku gdy kamery były ustawione na macierzy byłyby to widoki usytuowane najdalej od siebie, mające najmniej informacji wspólnej). W drugim atlasie przesyłane są wysokiej jakości głębie wejściowe odpowiadające widokom z atlasu pierwszego, które zostaną wykorzystane przez zmodyfikowane oprogramowanie IVDE po stronie dekodera. Geometria wyestymowana na podstawie wejściowych map głębi przesyłanych w drugim atlasie będzie miała lepszą jakość (co pokazano w podrozdziale 2.1). W dwóch pozostałych atlasach przesyłane są widoki wejściowe, które pozostały po dokonaniu selekcji do atlasu pierwszego. Zważywszy na dodanie dodatkowego atlasu map głębi, całkowita liczba przesyłanych widoków zostaje zmniejszona o liczbę widoków mieszczących się w jednym atlasie. Rys. 5 przedstawia jak wyglądają atlasy po wprowadzonych zmianach. Na Rys. 6 zaprezentowano jak wprowadzone zmiany wpłynęły na jakość zsyntezowanego w dekodерze widoku wirtualnego.



Rys. 5. Atlasy MIV DSDE po wprowadzonej modyfikacji. Każda kolumna widoków odpowiada jednemu atlasowi.



Rys. 6. Widoki wirtualne stworzone przy pomocy oprogramowania referencyjnego MIV DSDE: (A) i (B), oraz przy wprowadzonej modyfikacji: (A') i (B').

3. EKSPERYMENTY

3.1. Metodyka

Badania opisane w tym rozdziale zostały przedstawione w odniesieniu do wyników uzyskanych za pomocą MIV DSDE. Atlasy zawierające widoki wejściowe oraz widoki wejściowe z mapami głębi (w przypadku wprowadzonej modyfikacji) zostały zakodowane za pomocą jednej z wydajnych implementacji kodera VVC – Vvenc. Dane te zostały przesłane do dekodera, a następnie odczytane ze zdekodowanych strumieni bitów VVC. Następnie, zostały one wykorzystane do estymacji głębi przy użyciu oprogramowania IVDE. Testy zostały przeprowadzone na 15 różnych sekwencjach testowych, przy czym 9 tych sekwencji było wygenerowane komputerowo, a 6 z nich zawierało sceny naturalne. Sekwencje różniły się między sobą ułożeniem kamer oraz ich rodzajem (5 sekwencji pozyskanych kamerami wszechkierunkowymi, pozostałe 10 pozyskano zwykłymi kamerami perspektywicznymi). Sprawdzono w jaki sposób wprowadzona modyfikacja wpływa na prędkości bitowe oraz w jaki sposób zmieniły się czasy dekodowania.

3.2. Wyniki

Wyniki eksperymentów zostały podzielone na trzy podrozdziały. W podrozdziale 3.2.1. zaprezentowano jak zmienił się czas dekodowania dla każdej z sekwencji, w 3.2.2. pokazano zmianę przepływności przy zachowaniu stałej jakości, natomiast podrozdział 3.2.3. prezentuje ile średnio strumienia bitowego zajmuje przesyłana głębia w zmodyfikowanym TMIV.

3.2.1. Porównanie szybkości przetwarzania procesu kodowania i dekodowania

Wyniki pokazane w Tab. 1 ukazują czas działania metody wykorzystującej wejściowe mapy głębi wyrażony jako procent czasu działania metody referencyjnej. Kodowanie TMIV to proces selekcji widoków wejściowych oraz pakowania ich do atlasów, a kodowanie Vvenc to proces kompresji wizyjnej któremu poddawane są atlasy stworzone przez TMIV. Zwiększenie czasu kodowania jest efektem spodziewanym, jako iż zwiększono ilość danych wejściowych i wprowadzono dodatkowy krok identyfikowania widoków zawierających najwięcej informacji.

Tab. 1. Procentowa zmiana czasu obliczeń dla proponowanej metody w porównaniu do metody referencyjnej.

Sekwencja testowa	Kodowanie TMIV [%]	Kodowanie Vvenc [%]	Dekodowanie i synteza [%]
ClassroomVideo	1468,3	88,6	44,3
Museum	1901,6	122,1	65,6
Fan	367,9	114,0	63,5
Kitchen	804,2	101,5	59,1
Painter	358,6	113,9	59,4
Frog	285,4	101,1	43,2
Carpark	214,4	115,2	47,2
Chess	412,1	106,8	75,9
Group	558,3	85,6	58,9
Fencing	360,5	94,9	52,1
Hall	278,5	103,9	61,7
Street	439,4	130,7	57,7
ChessPieces	435,1	115,4	70,5
Hijack	785,7	96,1	56,6
Mirror	228,1	113,9	66,0
Cadillac	479,6	137,8	52,4

Warto jednak zaznaczyć, że udało się zmniejszyć czas dekodowania o połowę w porównaniu z metodą referencyjną. Jest to bardzo wskazane w przypadku MIV DSDE, jako że złożoność dekodera dla tego profilu jest bardzo duża, co jest wynikiem przeprowadzania w nim estymacji głębi.

3.2.2. Zmiana przepływności dla stałej jakości

W Tab. 2 pokazano wpływ zmiany przepływności dla stałej jakości sekwencji mierzonej IV-PSNR [5]. Na zielono zaznaczono te miejsca, w których przy zachowaniu tej samej jakości można zmniejszyć strumień skompresowanych danych (brak określenia procentowego oznacza, że krzywe jakości w zależności od przepływności nie nakładały się na siebie, dlatego niemożliwe było określenie procentowej zmiany przepływności). Jak można zauważyć, dla większości sekwencji następuje zmniejszenie przepływności, więc proponowane modyfikacje mogą zapewnić lepszą jakość wizji wszechogarniającej dostarczanej widzowi jeżeli zdecydujemy się na zachowanie tej samej przepływności.

Tab. 2. Zmiana przepływności w proponowanej metodzie w porównaniu z metodą referencyjną dla tej samej jakości mierzonej za pomocą IV-PSNR [5].

Sekwencja testowa	Niskie przepływności [%]	Wysokie przepływności [%]
ClassroomVideo	-43,2	-29,0
Museum	3,8	-4,0
Fan	-1,1	13,9
Kitchen	-14,3	-15,2
Painter	-37,4	-26,0
Frog	11,5	9,8
Carpark	-5,5	0,6
Chess	---	---
Group	---	---
Fencing	-24,8	-9,0
Hall	---	-82,4
Street	-6,5	-7,8
ChessPieces	---	---
Hijack	---	---
Mirror	7,2	-1,7
Cadillac	-19,8	-11,7

3.2.3. Średnia liczba bitów poświęconych na głębię

Wyniki zaprezentowane w Tab. 3 pokazują, że w zmodyfikowanej metodzie głębia zajmuje średnio jedynie 8% strumienia bitowego. Oznacza to, że jakość skompresowanych widoków wejściowych została jedynie nieznacznie zmniejszona, a jakość map głębi (jak pokazano na Rys. 4) została zauważalnie zwiększona. Daje to końcowy wzrost jakości widoku wirtualnego, co potwierdzają wyniki zaprezentowane w poprzednim podrozdziale.

Tab. 3. Podsumowanie całkowitej przepływności oraz zawartości strumienia bitowego.

Średnia przepływność [Mbps]		Część strumienia [%]					
MIV DSDE	Modyfikacja	MIV DSDE			Modyfikacja		
Całość strumienia (widoki + głębie + metadane)		Widoki	Metadane	Głębie	Widoki	Metadane	Głębie
36,870	35,751	100,0	0,0	0,0	95,8	0,0	4,2
21,332	20,620	99,9	0,1	0,0	94,2	0,1	5,7
11,644	11,209	99,9	0,1	0,0	91,9	0,1	8,0
5,970	5,815	99,8	0,2	0,0	89,4	0,2	10,4
3,039	2,900	99,6	0,4	0,0	86,8	0,4	12,8

4. PODSUMOWANIE

W tym artykule opisano usprawnienie procesu estymacji map głębi po stronie dekodera wizji wszechogarniającej. W metodzie referencyjnej do dekodera przesyłane

są tylko atlasy, w których znajdują się widoki zawierające najwięcej informacji o scenie. Estymacja głębi dokonywana jest po stronie dekodera tylko na podstawie widoków wejściowych.

W proponowanym podejściu do pierwszego z czterech atlasów zapisywane są widoki zawierające najwięcej informacji o scenie, a do drugiego atlasu pakowane są wysokiej jakości mapy głębi odpowiadające widokom z atlasu pierwszego. W pozostałych dwóch atlasach przesyłane są pozostałe widoki wejściowe.

Przeprowadzono wszechstronne eksperymenty by zaprezentować zalety wprowadzonych modyfikacji. Zbadano jak modyfikacje zmieniały przepływność bitową, w jaki sposób zmienia się przepływność przy zachowaniu stałej jakości obiektywnej wizji wszechogarniającej, oraz jaki procent strumienia bitowego stanowiła skompresowana wejściowa głębia. Wyniki eksperymentów ukazały, że udało się osiągnąć pożądaną poprawę przy zachowaniu takiej samej całościowej przepływności bitowej. Daje to możliwość uzyskania lepszej jakości przy zachowaniu stałej przepływności. Dodatkowo, udało się skrócić czas dekodowania (zawierający czas estymacji głębi) o średnio 50% w porównaniu do metody referencyjnej, znacząco zwiększając konkurencyjność metody kodowania wizji wszechogarniającej opartej na estymacji map głębi po stronie dekodera.

PODZIĘKOWANIA

Praca finansowana ze środków przyznanych przez Ministerstwo Edukacji i Nauki.

LITERATURA

- [1] Boyce J. M. i in. 2021. „MPEG Immersive Video Coding Standard”. *Proc. of the IEEE*.
- [2] Bross B. i in. 2021. „Overview of the Versatile Video Coding (VVC) standard and its applications”. *IEEE T Circ. & Sys. Vid. Tech.*
- [3] Domański M. i in. 2017. „Immersive visual media – MPEG-I: 360 video, virtual navigation and beyond”. *Proc. Int. Conf. Syst. Signals Image Process.*
- [4] Doumanoglou A. 2018. „Quality of experience for 3-D immersive media streaming”. *IEEE T Br 64*.
- [5] Dziembowski A. 2020. „Software manual of IV-PSNR software”.
- [6] ISO/IEC 23090-12. „Information technology - Coded representation of immersive media - Part 12: MPEG Immersive video”. ISO/IEC JTC 1/SC 29.
- [7] Mieloch D. i in. 2022. „Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video”. *IEEE T Circ. & Sys. Vid. Tech.*
- [8] Mieloch D. i in. 2020. „Depth Map Estimation for FreeViewpoint Television and Virtual Navigation”. *IEEE Access*.
- [9] Senoh T. i in. 2018. „Depth estimation and view synthesis for immersive media”. *IC3D conference*.
- [10] Sullivan G. i in. 2012. „Overview of the High Efficiency Video Coding (HEVC) standard”. *IEEE T Circ. & Syst. Vid. Tech.*
- [11] Tech G. i in. 2016. „Overview of the Multiview and 3D Extensions of High Efficiency Video Coding”. *IEEE T Circ. & Sys. Vid. Tech.*
- [12] Wien M. i in. 2019. „Standardization Status of Immersive Video Coding”. *IEEE J Em. & Sel. Top. Circ. & Sys.*