Adrian Dziembowski
Adam Grzelka
Dawid Mieloch
Katedra Telekomunikacji Multimedialnej i Mikroelektroniki
Politechnika Poznańska
ul. Polanka 3, 60-965 Poznań
{adziembowski, agrzelka, kwegner}@multimedia.edu.pl

Kraków, 27-29 czerwca 2016

WIELOWIDOKOWA SYNTEZA W SYSTEMACH TELEWIZJI SWOBODNEGO PUNKTU WIDZENIA

MULTIVIEW SYNTHESIS IN FREE-VIEWPOINT TELEVISION SYSTEMS

Streszczenie**: W artykule opisano metodę syntezy widoków wirtualnych wykorzystującą informację o teksturze i głębi z wielu widoków rzeczywistych. Przedstawione wyniki eksperymentalne potwierdzają znaczącą poprawę jakości syntezy w porównaniu do algorytmów wykorzystujących jedynie dwa sąsiednie widoki.**

Abstract**: In the article we described a method of virtual view synthesis, which uses texture and depth information from multiple real views. Presented results confirm significant quality improvement compared to algorithms based on two neighboring real views only.**

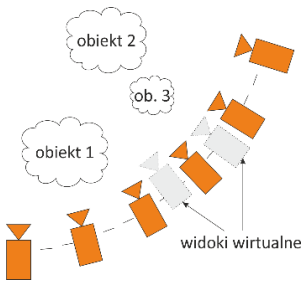
Słowa kluczowe**: synteza widoków wirtualnych, telewizja swobodnego punktu widzenia, swobodna nawigacja, system wielokamerowy**

Keywords**: virtual view synthesis, free-viewpoint television, free navigation, multicamera system**

1. WSTĘP

Telewizja swobodnego punktu widzenia (ang. Free-Viewpoint Television – FTV) jest usługą pozwalającą widzowi na swobodną nawigację wokół sceny zarejestrowanej przy użyciu odpowiedniego systemu wielokamerowego. Systemy tego rodzaju mogą zawierać nawet kilkaset kamer [10], jednakże powszechniejsze i zdecydowanie praktyczniejsze są systemy liczące kilka do kilkunastu kamer [4],[5].

W celu zapewnienia w pełni swobodnej nawigacji użytkownika, wybór jego punktu widzenia, czyli punktu, z którego obserwuje on scenę, nie może ograniczać się wyłącznie do fizycznych pozycji kamer (zwłaszcza w systemach, w których odległości pomiędzy kamerami są znaczne).



Rys. 1. Idea systemu telewizji swobodnego punktu widzenia

Aby umożliwić płynny ruch widza wokół sceny, konieczne jest stworzenie (synteza) pośrednich, wirtualnych widoków – obrazów, które byłyby zarejestrowane przez kamery stojące pomiędzy rzeczywistymi urządzeniami. Ideę tę pokazano na rys. 1.

Widoki wirtualne w systemach telewizji swobodnego punktu widzenia tworzone są na podstawie przestrzennego modelu sceny, zazwyczaj danego w postaci estymowanych na podstawie rzeczywistych widoków map głębi.

2. SYNTEZA WIDOKÓW WIRTUALNYCH

2.1. Przetwarzanie wstępne

Będąca przedmiotem artykułu synteza widoków wirtualnych stanowi ostatni etap przetwarzania danych wizyjnych w systemie telewizji swobodnego punktu widzenia. Jednakże, aby wygenerowanie pośrednich obrazów było możliwe, konieczne jest wcześniejsze przetworzenie zarejestrowanych obrazów.

Pierwszym krokiem prowadzącym do umożliwienia widzowi swobodnej nawigacji jest kalibracja systemu wielokamerowego, którym dokonywana jest akwizycja materiału. Na tym etapie wyznaczana jest:

- charakterystyka optyczna każdej kamery (parametry wewnętrzne, ang. intrinsic parameters) wraz z parametrami zniekształceń wprowadzanych przez niedoskonałość soczewek (ang. lens distortions), estymowana oddzielnie dla każdej kamery,
- względne rozmieszczenie i rotacja wszystkich kamer systemu (parametry zewnętrzne, ang. extrinsic parameters) [6].

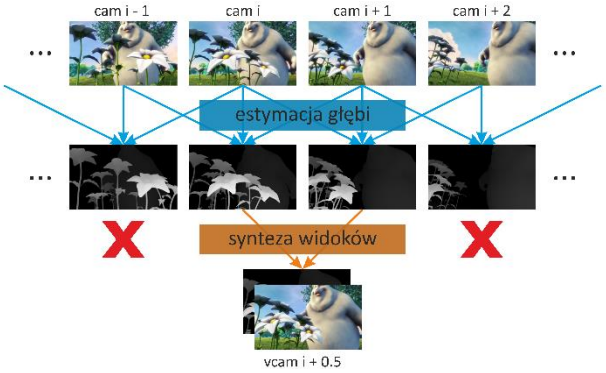
Po dokonaniu kalibracji systemu, a przed procesem syntezy, konieczne jest wyznaczenie przestrzennego modelu zarejestrowanej sceny. W tym celu przeprowadzany jest proces estymacji głębi, czyli informacji o odległości dzielącej każdy zarejestrowany obiekt od poszczególnych kamer. Informacja ta jest zazwyczaj zapisywana w postaci map głębi – odpowiadających obrazom zarejestrowanym przez kamery obrazów, na których zapisana jest odległość każdego widocznego punktu (im jaśniejszy jest dany punkt na mapie głębi, tym bliżej kamery się on znajduje).

W celu osiągnięcia dużej dokładności, oba opisane procesy powinny być przeprowadzone jednocześnie dla

wszystkich kamer wchodzących w skład systemu. W pierwszym etapie zwiększa to dokładność estymowanych parametrów. W przypadku estymacji głębi, wykorzystywanie wielu widoków jednocześnie wiąże się z dwiema podstawowymi zaletami: możliwością wyznaczenia głębi w obszarach przysłoniętych w części kamer oraz zmniejszeniem wpływu szumu w obrazach wejściowych na poprawną estymację głębi.

2.2. Synteza z dwóch widoków

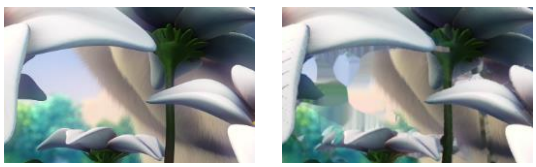
W odróżnieniu od wcześniejszych etapów przetwarzania obrazu w systemie wielowidokowym, gdzie do estymacji – czy to parametrów zewnętrznych, czy też map głębi – wykorzystywana była jednocześnie informacja ze wszystkich kamer, najpopularniejsze podejście do syntezy widoków wirtualnych uwzględnia wykorzystanie jedynie dwóch sąsiednich widoków rzeczywistych (rys. 2). Taki algorytm używany jest chociażby w stworzonym na potrzeby grupy MPEG oprogramowaniu modelowym do syntezy widoków wirtualnych – VSRS (View Synthesis Reference Software) [9].



Rys. 2. Algorytm syntezy z dwóch widoków rzeczywistych

Na rys. 2 w sposób schematyczny pokazano najpopularniejszy sposób syntezy widoku wirtualnego: do syntezy widoku $i + 0.5$ wykorzystywana jest informacja o głębi i teksturze wyłącznie z widoków i oraz $i + 1$, pomimo faktu, iż dostępne są dalsze lewe i prawe widoki ($i – 1$, $i + 2$ i kolejne).

W takim podejściu, w widoku wirtualnym mogą występować rozległe obszary odsłonięte – tym większe, im większa jest dynamika głębi zarejestrowanej sceny. Obszary te reprezentują te regiony sceny, które były przysłonięte zarówno w lewym, jak i w prawym widoku rzeczywistym.

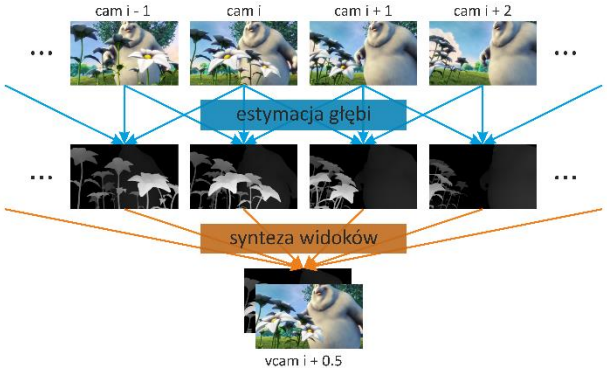


Rys. 3. Porównanie fragmentu oryginalnego widoku z sekwencji BBB Flowers z obrazem zsyntezowanym (przy użyciu VSRS) z wypełnionymi odsłonięciami

W przypadku syntezy z dwóch tylko widoków, w obszarach tych nie jest znana ani głębia, ani tekstura, której to znajomość jest konieczna z punktu widzenia końcowego użytkownika systemu swobodnej nawigacji. Odsłonięte regiony widoku wirtualnego muszą więc zostać wypełnione (ang. inpainting) na podstawie najbliższego otoczenia, co przy skomplikowanej teksturze przysłanianych obiektów przekłada się na spadek subiektywnej jakości syntezowanego widoku (rys. 3).

2.3. Synteza wielowidokowa

Rozwiązaniem opisanych w poprzednim punkcie problemów jest zastosowanie algorytmu syntezy wykorzystującego więcej widoków rzeczywistych, aniżeli tylko dwa sąsiednie (rys. 4). Podejście takie nie jest rozwiązaniem nowym [2],[7].



Rys. 4. Algorytm syntezy wielowidokowej

W syntezie wielowidokowej, zmarginalizowany jest wpływ przysłonięć – kiedy scena rejestrowana jest przez większą liczbę kamer, obszarów całkowicie przysłoniętych we wszystkich widokach rzeczywistych jest zdecydowanie mniej, niż w przypadku syntezy dwuwidokowej. Z tego powodu, na znacznie mniejszym obszarze dowolnego widoku wirtualnego istnieje konieczność algorytmicznego jego wypełnienia.

Ponadto, w syntezie wielowidokowej wszystkie widoki wirtualne syntezowane są przy wykorzystaniu tej samej informacji wejściowej (zestawu wszystkich widoków rzeczywistych wraz z odpowiadającymi im mapami głębi). Zapewnia to zdecydowanie większą spójność międzywidokową, niż w syntezie z dwóch widoków, co skutkuje większą subiektywnie jakością obrazu podczas swobodnej nawigacji końcowego użytkownika systemu.

Niemniej jednak, synteza wielowidokowa nie jest pozbawiona wad. Po pierwsze, charakterystyka barwna wszystkich widoków jest uśredniana, co w systemie wielokamerowym o bardzo szerokim kącie widzenia powoduje wrażenie nienaturalności.

Ponadto, gdy osie optyczne kamery wirtualnej i rzeczywistej skierowane są pod znacznym kątem (bliskim 90°), skończona reprezentacja map głębi (zazwyczaj przesyłanych w postaci 8-bitowej) może skutkować znaczącymi błędami przy rzutowaniu punktów do widoku wirtualnego. Korzystanie z danego widoku rzeczywistego przy syntezie skutkować więc będzie znaczącymi przekłamaniami w okolicach krawędzi obiektów na widoku wirtualnym.

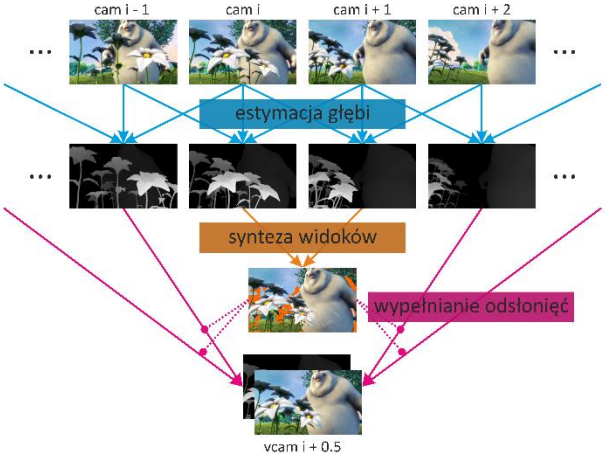
Dużą wadą syntezy z wielu widoków jest również większa złożoność obliczeniowa takiego algorytmu – w przypadku dziesięciu kamer jest ona w przybliżeniu

pięciokrotnie większa, niż dla syntezy dwuwidokowej. Fakt ten ma duże znaczenie w projektowaniu systemu telewizji swobodnego punktu widzenia, gdzie w celu umożliwienia swobodnej nawigacji widza wewnątrz sceny operacja syntezy musi być przeprowadzana w czasie rzeczywistym.

3. PROPONOWANA METODA SYNTEZY

3.1. Idea

Rozważywszy wady i zalety obu opisywanych metod syntezy, zdecydowano się stworzyć metodę łączącą oba podejścia. Zaproponowano, aby synteza widoku wirtualnego przebiegała dwuetapowo. W etapie pierwszym, widok pośredni syntezowany jest wyłącznie z dwóch sąsiednich widoków rzeczywistych – tak, jak w oprogramowaniu modelowym. W kolejnym kroku jednak (w odróżnieniu od VSRS, gdzie obszary niezsyntezowane wypełniane były na podstawie najbliższego otoczenia), obszary odsłonięte są wypełniane informacją przerzutowaną z pozostałych, dalszych widoków rzeczywistych (rys. 5).



Rys. 5. Algorytm syntezy dwuwidokowej z wielowidokowym wypełnianiem odsłonięć

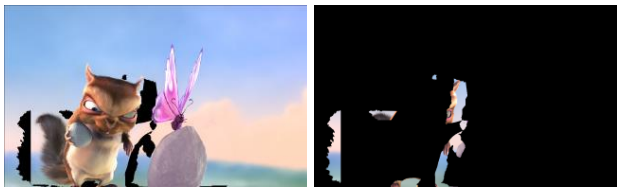
Zaproponowana metoda znacząco ogranicza obecny w syntezie dwuwidokowej problem wypełniania obszarów odsłoniętych. Jednocześnie, zachowuje ona podstawowe zalety tego rodzaju syntezy, takie jak krótki czas obliczeń oraz zachowanie naturalnej charakterystyki barwnej syntezowanego widoku.

3.2. Opis algorytmu

Pierwszym krokiem wiodącym do zsyntezowania widoku wirtualnego jest przerzutowanie informacji z dwóch sąsiednich widoków rzeczywistych. W tym celu zastosowano syntezę „w tył" (ang. backward synthesis) [3]. W tym podejściu, najpierw tworzona jest wirtualna mapa głębi – do tworzonego widoku rzutowana jest wyłącznie informacja o przestrzennym rozmieszczeniu obiektów, a pomijana jest tekstura. Następnie, stworzona i odpowiednio przefiltrowana (w celu usunięcia dziur i artefaktów wynikających ze skończonej rozdzielczości obrazu i map głębi) wirtualna mapa głębi jest wykorzystywana do przerzutowania informacji o teksturze z widoków rzeczywistych do tworzonego widoku pośredniego. Jeżeli dany punkt widoku wirtualnego widoczny był tylko w jednym z rzeczywistych obrazów, jego barwa jest kopiowana, w przeciwnym wypadku – z kolorów odpowiadających punktów w obu widokach rzeczywistych liczona jest średnia ważona, gdzie wagami są: odległość pomiędzy wirtualnym i rzeczywistym widokiem oraz głębia odpowiadającego punktu w obrazie rzeczywistym.

W drugim etapie syntezy dokonywane jest wypełnianie obszarów niezsyntezowanych w etapie pierwszym. Aby zachować jak największe podobieństwo charakterystyki tychże obszarów z pozostałą częścią obrazu, do ich wypełnienia wykorzystywana jest informacja z możliwie najbliższych kamer. Dla przykładu z rys. 5, obszary niezsyntezowane w widoku $i + 0.5$ w pierwszej kolejności zostaną wypełnione teksturą z widoków $i - 1$ oraz $i + 2$. Dopiero w przypadku, gdy widok $i + 0.5$ wciąż będzie zawierał obszary bez informacji, te zostaną przerzutowane z dalszych kamer.

By zmniejszyć czas obliczeń, do widoku wirtualnego rzutowane są wyłącznie te fragmenty obrazów rzeczywistych, na których wystąpiły odsłonięcia. Wybór tych obszarów jest dokonywany poprzez przerzutowanie otoczenia niezsyntezowanych obszarów z widoku wirtualnego do rzeczywistego. Następnie, w drugą stronę rzutowane są wyłącznie obszary znajdujące się wewnątrz tego przerzutowanego obszaru (rys. 6). Dla każdego kolejnego analizowanego widoku rzeczywistego jest coraz mniej niezsyntezowanych regionów, a więc liczba punktów, które trzeba przerzutować maleje. W ogólności, średni zmierzony narzut obliczeń dla operacji wypełniania odsłonięć w widoku wirtualnym wynosi niecałe 20% czasu syntezy dwuwidokowej.



Rys. 6. Wypełnianie odsłonięć (sekw. BBB Butterfly): z lewej: widok zsyntezowany z dwóch sąsiednich kamer, po prawej: analizowane obszary z kolejnego widoku

Tak jak w pierwszym etapie, również przy wypełnianiu odsłonięć wykorzystywana jest synteza „w tył", pozwalająca na filtrację informacji o głębi przed rzutowaniem tekstury.

Ostatnim krokiem procesu syntezy widoku wirtualnego jest wypełnienie obszarów niewidocznych w żadnej z rzeczywistych kamer. Obszary te są interpolowane na podstawie najbliższego otoczenia [1]. Niemniej jednak, procentowy udział takich obszarów w całym obrazie jest zdecydowanie mniejszy, niż w przypadku zwykłej syntezy dwuwidokowej.

4. WYNIKI EKSPERYMENTALNE

Do przeprowadzenia testów eksperymentalnych zdecydowano się wykorzystać zbiór sześciu syntetycznych sekwencji testowych: Big Buck Bunny [8] Butterfly, Flowers i Rabbit – wszystkie zarówno dla liniowego, jak i łukowego rozmieszczenia kamer.

Rys. 7. Przykładowe widoki dla sekwencji (od lewej): BBB Butterfly, BBB Flowers i BBB Rabbit

Pierwszym powodem odrzucenia sekwencji rzeczywistych był brak widoków odniesienia (zarejestrowanych w miejscach syntezowanych widoków wirtualnych), co skutkowałoby niemożnością estymacji obiektywnej jakości syntezowanych widoków wirtualnych. Po drugie, dla wybranych sekwencji syntetycznych dostępne są idealne mapy głębi, dzięki czemu ocenie poddana zostanie wyłącznie operacja syntezy, a nie synteza wraz z estymacją głębi.

Dla każdej z użytych sekwencji dostępne jest 91 widoków. Do testów zdecydowano się użyć 79 z nich (od 6 do 84 włącznie). Siedem z nich było traktowane jako widoki rzeczywiste (kolejno widoki: 6, 19, 32, 45, 58, 71 i 84), pozostałe stanowią jedynie odniesienie do estymacji jakości syntezy.

Do obiektywnej oceny zaproponowanego algorytmu syntezy zdecydowano się użyć miary PSNR. W tym celu, dla każdej analizowanej sekwencji syntezowano 72 widoki wirtualne – odpowiadające pozycjom wszystkich widoków odniesienia. Następnie, w każdej z tych pozycji mierzono PSNR pomiędzy oryginalnym obrazem a obrazem zsyntezowanym.

W celu porównania wyników, identyczną procedurę oceny jakości wykonano zarówno dla zaproponowanej metody syntezy, jak i dla syntezy dwuwidokowej wykorzystywanej w oprogramowaniu modelowym.

W tabeli 1 przedstawiono uśrednione dla wszystkich widoków wyniki dla każdej z użytych sekwencji.

Tab. 1. Porównanie jakości syntezy w oprogramowaniu modelowym z zaproponowaną metodą

Sekwencja	*PSNR [dB]*		*Zysk*
	VSRS	*propozycja*	*[dB]*
BBB Butterfly Linear	37,67	39,41	2,30
BBB Butterfly Arc	36,37	38,39	2,02
BBB Flowers Linear	28,43	31,30	2,87
BBB Flowers Arc	26,77	29,95	3,18
BBB Rabbit Linear	36,13	38,12	1,99
BBB Rabbit Arc	31,38	33,73	2,35
średnio	32,79	35,15	2,36

Jak pokazano, zaproponowana metoda pozwala na osiągnięcie lepszej jakości dla wszystkich testowanych sekwencji. Średnio, metoda z wielowidokowym wypełnianiem odsłonięć pozwala zwiększyć jakość o 2,36 dB.

Również subiektywna jakość widoków zsyntezowanych przy użyciu syntezy dwuwidokowej z wielowidokowym wypełnianiem odsłonięć jest lepsza, niż dla metody odniesienia. Na rys. 8 przedstawiono porównanie fragmentu widoku odniesienia z sekwencji BBB Flowers i widoku zsyntezowanego przy użyciu proponowanej metody. Ten sam fragment zsyntezowany przy użyciu VSRS przedstawiono na rys. 3.

Rys. 8. Porównanie fragmentu oryginalnego obrazu i obrazu syntezowanego proponowaną metodą

5. PODSUMOWANIE

W artykule opisano metodę syntezy widoków wirtualnych, mogącą znaleźć zastosowanie w systemach telewizji swobodnego punktu widzenia. W zaproponowanej metodzie, większość widoku wirtualnego syntezowana jest przy użyciu informacji z dwóch sąsiednich widoków. Obszary niewidoczne na obu sąsiednich widokach rzeczywistych są natomiast rzutowane z widoków dalszych.

Zaproponowane podejście pozwala na osiągnięcie jakości syntezy lepszej o ponad 2,3dB niż przy użyciu oprogramowania modelowego, przy nieznacznym tylko zwiększeniu nakładu obliczeń.

PODZIĘKOWANIA

Praca sfinansowana ze środków Narodowego Centrum Badań i Rozwoju w ramach Umowy nr TANGO1/266710/NCBR/2015.

LITERATURA

[1] Bertalmio M., G. Sapiro, V. Caselles, C. Ballester. 2000. „Image inpainting". *Computer Graphics*.

[2] Cooke E., P. Kauff, T. Sikora. 2006. „Multi-view synthesis: A novel view creation approach for free viewpoint video". *Signal Processing: Image Communication, 21*, 476-492.

[3] Du-Hsiu L., H. Hsueh-Ming, L. Yu-Lun. 2013. „Virtual view synthesis using backward depth warping algorithm". *Picture Coding Symposium, PCS*.

[4] Dziembowski A., A. Kuehn, A. Łuczak, D. Mieloch, K. Wegner. 2014. „Realizacja eksperymentalnego systemu telewizji swobodnego punktu widzenia z łukowym ustawieniem kamer". *Przegląd Telekomunikacyjny, 6/2014*, 161-164.

[5] Goorts P., L. Jorissen, G. Lafruit. 2014. "[FTV AHG] EE2 results on Soccer-Linear1 with foreground/background segmentation". *ISO/IEC JTC1/SC29/WG11, Doc. M34310*.

[6] Heyden A, M. Pollefeys. 2004. „Multiple view geometry". *Emerging topics in computer vision*, 63-75.

[7] Jun-Te H., L. Jin-Jang. 2015. „Virtual view synthesis for multi-view video plus depth sequences using spatial-temporal information". *3DTV-CON*.

[8] Kovacs, P.. 2015. "[FTV AHG] BBB light-field test sequences". *ISO/IEC JTC1/SC29/WG11, M35721*.

[9] Stankiewicz O., K. Wegner, M. Tanimoto, M. Domański. 2013. „Enhanced view synthesis reference software (VSRS) for Free-viewpoint Television". *ISO/IEC JTC1/SC29/WG11, Doc. M31520*.

[10] Tanimoto M.. 2010. „FTV (Free-Viewpoint TV)". *Proc. of 2010 IEEE 17th Int. Conf. on Image Proc*.