KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2024

TWORZENIE WIELOWIDOKOWYCH SEKWENCJI DLA WIZJI WSZECHOGARNIAJĄCEJ

CREATING MULTIVIEW SEQUENCES FOR IMMERSIVE VIDEO

Dawid Mieloch; Dominika Klóska; Adrian Dziembowski; Błażej Szydełko; Adam Grzelka; Jakub Stankowski

Instytut Telekomunikacji Multimedialnej, Politechnika Poznańska, Poznań
{dawid.mieloch,dominika.kloska,adrian.dziembowski,blazej.szydelko,adam.grzelka,jakub.stankowski}@put.poznan.pl

Streszczenie**: Wielowidokowe sekwencje wizyjne są niezbędne do rozwoju technologii wykorzystywanych dla celów wizji wszechogarniającej. Szczególnie ważną częścią tych badań są prace nad nowymi normami kompresji. W tym artykule przedstawiono przegląd metod, których użycie w procesie tworzenia sekwencji ma korzystny wpływ na ich końcową jakość, co potwierdzają wyniki przeprowadzonych badań eksperymentalnych. Proces tworzenia wielowidokowej sekwencji został przedstawiony na przykładzie sekwencji Choreo, obecnie wykorzystywanej w pracach ISO/IEC MPEG Implicit Neural Visual Representation.**

Abstract**: Multiview sequences are essential to develop technologies utilized for immersive video purposes. Especially important part of this research involves work on the new compression standards. This article provides an overview of methods that positively influence the final result of the test sequence creation process, as confirmed by results of conducted experiments. The process of multiview sequence acquisition was presented using the example of the Choreo sequence which is now used in the work of ISO/IEC MPEG Implicit Neural Visual Representation.**

Słowa kluczowe**: system wielokamerowy. wizja wszechogarniająca.**

Keywords**: multicamera system, immersive video.**

1. WSTĘP

Wizja wszechogarniająca (immersyjna) jest tematem cieszącym się dużym zainteresowaniem ze strony mediów społecznościowych [9], branży turystycznej [8], edukacyjnej [24] i wielu innych. Użytkownik systemu wykorzystującego wizję wszechogarniającą może oglądać zarejestrowaną scenę z dowolnego punktu widzenia i nie musi ograniczać się do położeń, w których usytuowano wcześniej kamery. Taka możliwość zapewniona jest dzięki zareprezentowaniu całej trójwymiarowej sceny. Typowo, wizja wszechogarniająca wykorzystuje format MVD (Multiview Video + Depth) [18], na który składają się widoki zarejestrowane z różnych punktów widzenia oraz odpowiadające im mapy głębi, które zawierają informacje o geometrii sceny (Rys. 1).

W systemach wielokamerowych, które służą do rejestracji wizji wszechogarniającej, można stosować dowolne ułożenie kamer, zarówno perspektywicznych jak i dookólnych [2]. Największym ograniczeniem jest wymóg, aby kamery posiadały wejście synchronizacyjne, co znacznie ogranicza liczbę modeli które można z powodzeniem wykorzystać w takim systemie [5].

Rys.1 Sekwencja wielowidokowa Choreo [13], zarejestrowane widoki i odpowiadające im mapy głębi.

Proces pozyskiwania informacji o geometrii sceny różni się w zależności od typu sekwencji. W przypadku sekwencji generowanych komputerowo, geometria sceny zazwyczaj może być wyeksportowana z programu użytego do ich stworzenia (jest to możliwe np. w powszechnie używanym programie Blender [23]). Informację o geometrii sceny sekwencji naturalnej można pozyskać wykorzystując kamery głębi, które najczęściej mierzą odległość do obiektów znajdujących się w scenie za pomocą promieni podczerwonych [12]. Rozwiązanie to ma swoje wady, m.in. ograniczoną rozdzielczość nagranych map głębi czy możliwe interferencje wzbudzane przez sygnał sąsiadujących ze sobą kamer [26]. W związku z tym, w celu pozyskania informacji o geometrii sceny, najczęściej wykorzystywana jest estymacja głębi, przeprowadzana na podstawie parametrów kamer (czyli ich parametrów optycznych oraz ich położenia) oraz widoków nagranych przez system wielokamerowy [17].

Rosnąca popularność wizji immersyjnej spowodowała zwiększenie liczby badań prowadzonych w ramach rozwoju tej technologii ze szczególnym uwzględnieniem zagadnień kompresji [25] – pojawiła się pierwsza norma kompresji takiej wizji, czyli MPEG Immersive Video [2]. Ilość materiałów testowych stworzonych na potrzeby badawcze wizji wszechogarniającej jest niestety stosunkowo niewielka w porównaniu z ogromem danych testowych, wykorzystywanych na przykład do trenowania sieci neuronowych. Stworzenie wielowidokowej sekwencji, która może zostać wykorzystana do testów nie należy jednak do łatwych zadań, szczególnie w przypadku sekwencji naturalnych. Pod uwagę należy wtedy wziąć wiele czynników, które mogą negatywnie wpłynąć na końcową jakość sekwencji jeżeli nie zostaną one odpowiednio zaadresowane.

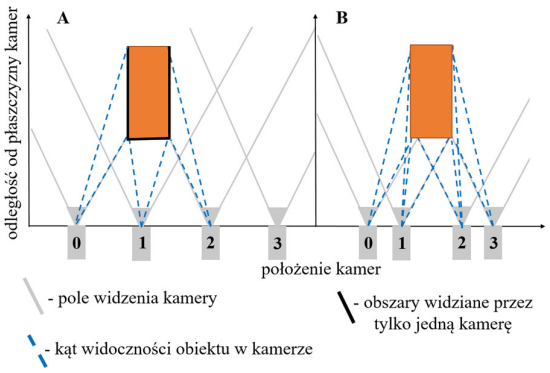
Aby ułatwić proces tworzenia sekwencji testowych, w niniejszym artykule przedstawiono przegląd metod, których zastosowanie zalecane jest w procesie akwizycji wielowidokowej sekwencji na potrzeby wizji wszechogarniającej. W kolejnym rozdziale opisano kolejne etapy rejestracji i przetwarzania sekwencji na przykładzie procesu akwizycji sekwencji testowej Choreo [13]. W rozdziale trzecim zaprezentowano wyniki eksperymentalne obrazujące wpływ poprawności przeprowadzenia kolejnych etapów tworzenia sekwencji na jej końcową jakość, a co za tym idzie, jej użyteczność w badaniach.

2. REJESTRACJA SEKWENCJI

2.1. Wybór kamer i ich rozstawienie

Proces nagrywania sekwencji testowej należy rozpocząć od ustalenia ułożenia oraz liczby kamer w systemie wielokamerowym. Aby możliwe było wyznaczenie głębi punktu w scenie konieczne jest użycie minimum dwóch kamer, jednak zwiększenie liczby kamer w systemie będzie miało pozytywny wpływ na jakość estymacji głębi [17]. Należy podkreślić, że wszystkie kamery muszą być ze sobą zsynchronizowane aby akwizycja w każdej z nich rozpoczynała się w tym samym momencie [20].

Najczęściej spotykanym sposobem rozstawienia kamer jest umieszczenie ich w równomiernych odstępach [5]. Odległość między poszczególnymi kamerami musi być jednak tak dopasowana, by nie dopuścić do sytuacji w której dany obiekt jest widziany tylko przez jedną kamerę, ponieważ wtedy estymacja głębi tego obiektu będzie niemożliwa (Rys. 2 A).



Rys. 2 (A) System wielowidokowy z niepoprawnym rozstawieniem kamer. (B) Ten sam system po zmniejszeniu odległości i zastosowaniu grupowania w pary.

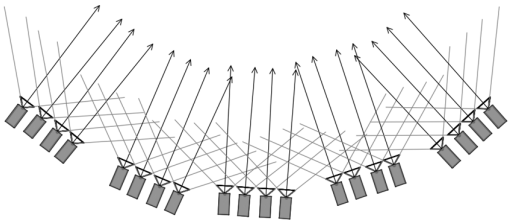
Jeżeli scena zawiera przesłonięte elementy, znalezienie odpowiedniej odległości między kamerami w rozstawieniu równomiernym może okazać się niemożliwe. Odpowiedzią na powyższe problemy jest łączenie kamer w równomiernie rozstawione grupy [20] (Rys. 2 B). Takie też rozstawienie zastosowano przy tworzeniu sekwencji Choreo. Rys. 3 obrazuje dokładne ułożenie systemu wielokamerowego podczas akwizycji tej sekwencji.

2.2. Kalibracja systemu wielokamerowego

Nieodłącznym elementem procesu tworzenia sekwencji wielowidokowej jest pozyskanie danych o położeniu i ustawieniu każdej z kamer w scenie. Aby pozyskać te dane należy przeprowadzić kalibrację systemu wielokamerowego, która obejmuje zarówno wyznaczenie parametrów wewnętrznych kamer, jak i relacji przestrzennych między nimi, czyli kalibracji zewnętrznej.

A)

B)



Rys. 3 System wielokamerowy wykorzystany do akwizycji sekwencji Choreo. A) System, widoczny po prawej stronie nagrywanej sceny. B) Wizualizacja rozstawienia kamer, strzałki symbolizują osie optyczne.

Parametry wewnętrzne opisują układ optyczny kamery, czyli długość ogniskowej, współrzędne punktu środkowego matrycy i współczynniki zniekształceń soczewkowych. Parametry zewnętrzne definiują natomiast współrzędne położenia kamery w trzech płaszczyznach oraz wartości kąta obrotu wokół każdej osi, czyli kierunek w którym zwrócona jest oś optyczna kamery [21].

Bez względu na rodzaj parametrów, problem kalibracji można w ogólności można rozłożyć na dwa etapy. Pierwszy etap polega na zebraniu danych odniesienia poprzez ekstrakcję punktów charakterystycznych z zarejestrowanych przez kamery widoków. W drugim etapie uzyskane dane stanowią podstawę do rozwiązania problemu optymalizacji funkcji. Proces ten polega na minimalizacji różnicy między wyliczanymi a rzeczywistymi pozycjami punktów charakterystycznych w obrazie [11].

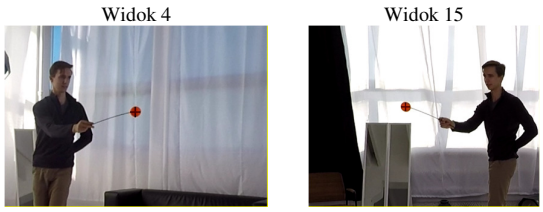
Najprostszą i najbardziej niezawodną metodą kalibracji parametrów wewnętrznych jest metoda z użyciem znacznika w formie czarno-białej szachownicy [28], przeprowadzana niezależnie dla każdej kamery. Istotnym aspektem tej metody jest to, że szachownica musi być płaska i sztywna co zapewnia dokładność detekcji punktów charakterystycznych. Kalibracja wewnętrzna nie musi odbywać się w miejscu akwizycji sekwencji, co umożliwia przeprowadzenie jej w bardziej dogodnych warunkach oświetleniowych i przy większej kontroli pozycjonowania znacznika w kadrze.

Większą uwagę należy zwrócić na kalibrację parametrów zewnętrznych, która przeprowadzana na miejscu rejestracji sekwencji i obejmuje wszystkie kamery. Klasyczne metody wyznaczania parametrów zewnętrznych obejmują kalibrację z wykorzystaniem punktów charakterystycznych otrzymanych za pomocą tablicy kalibracyjnej ze wzorem. Są to przede wszystkim: omawiana przy kalibracji wewnętrznej czarno-biała szachownica [21] lub

znaczniki ArUco [19]. Należy jednak zaznaczyć, że wzorce wykorzystywane w tych metodach powinny być dobrze widoczne z każdej perspektywy i w różnych orientacjach, co w przypadku dowolnego ustawienia kamer może być niemożliwe do osiągnięcia.

Odmiennym podejściem charakteryzują się metody bezznacznikowe, określane również jako automatyczne [1]. Bazują one na ekstrakcji punktów charakterystycznych z lokalnych cech obrazu, np. krawędzi i narożników. Popularnym algorytmem stosowanym do ekstrakcji cech i wyznaczania relacji między obrazami jest SIFT (ang. *Scale-Invariant Feature Transform*), jednak zazwyczaj jest on stosowany do kalibracji stereopar [15]. Bardziej zaawansowane techniki autokalibracji, takie jak [4], [16], charakteryzują się estymacją parametrów zewnętrznych na podstawie ruchu kamer, co wyklucza ich stosowanie przy tworzeniu sekwencji wielowidokowej na potrzeby wizji wszechogarniającej.

Jednym z podejść, które zapewnia wymaganą praktyczność i wygodę stosowania jest metoda wykorzystująca prosty znacznik kalibracyjny w formie kolorowej kuli [10], [23]. Taki znacznik jest prosty w obsłudze, a jego kształt sprawia, że z każdej perspektywy jest widziany tak samo i łatwo znaleźć jego położenie (Rys. 4).



Rys. 4 Wyznaczanie parametrów zewnętrznych sekwencji Choreo z użyciem prostego znacznika.

2.3. Estymacja głębi

Najczęstsze wymagania stawiane metodom estymacji głębi dla systemów wizji wszechogarniającej to estymacja spójna międzykamerowo oraz międzyramkowo, możliwa do przeprowadzenia dla różnego typu kamer (w tym dookólnych) [17].

Do najważniejszych metod estymacji głębi należą metody klasyczne, działające najczęściej poprzez minimalizację funkcji celu, a najczęściej wykorzystywany algorytm cięcia grafu (ang. *Graph Cut*) [14]. Drugą silnie rozwijaną w ostatnich latach grupą metod są algorytmy wykorzystujące uczenie maszynowe [27], [29]. Niestety, zapotrzebowanie na pamięć oraz moc obliczeniową w takich algorytmach zwiększa się wraz z liczbą widoków wejściowych, co uniemożliwia estymację map głębi w pełnej rozdzielczości.

Dla sekwencji Choreo, oraz szeregu innych testowych sekwencji wielowidokowych, została wykorzystana klasyczna metoda rozwijana przez ekspertów ISO/IEC MPEG Video Coding – Immersive Video Depth Estimation (IVDE) [17]. Oprogramowanie to wykorzystuje metodę segmentacji widoków wejściowych, optymalizację przez algorytm cięcia grafu, oraz szereg narzędzi zapewniających spójność czasową i zmniejszających złożoność obliczeniową. Działa dla dowolnego rodzaju oraz liczby kamer, przy niedużym zapotrzebowaniu na pamięć.

2.4. Dodatkowe przetwarzanie

Estymacja map głębi może być poprzedzona poprzez dodatkowe kroki, które mogą zapewnić wyższą jakość wizji dostarczanej końcowemu widzowi, takie jak międzywidokowa korekcja barwna czy odszumienie wejściowych sekwencji.

Należy zaznaczyć, że kroki te nie są niezbędne, jednak pozwalają podnieść jakość stworzonej sekwencji. Przeprowadzenie korekcji barwnej znacząco zmniejsza wpływ różnych charakterystyk barwnych używanych kamer na końcową jakość widoku wirtualnego [6], co potwierdzono również w jednym z przeprowadzonych w rozdziale trzecim badań. Usunięcie szumu znacząco wpływa na poprawność międzywidokowego pasowania bloków w estymacji głębi i na jej spójność w czasie [21].

3. EKSPERYMENT

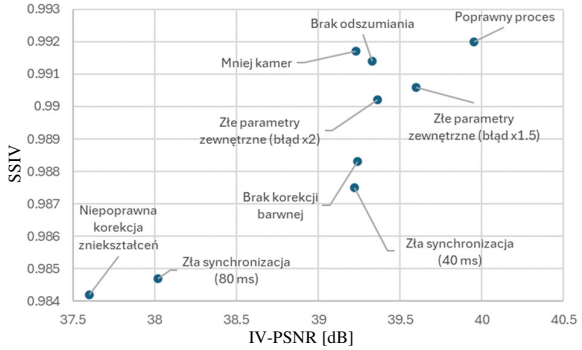
W celu zaprezentowania wpływu poszczególnych etapów przetwarzania sekwencji wielowidokowej na finalną jakość wizji wszechogarniającej, przeprowadzono eksperyment, w którym modyfikowano proces tworzenia takiej sekwencji. Badanie przeprowadzono poprzez pomijanie wykorzystywania poszczególnych narzędzi, lub też ich niepoprawne użycie.

Eksperyment przeprowadzony został dla sekwencji Choreo, dla której, w każdej testowanej konfiguracji estymowano mapy głębi dla widoków *v6, v7, v8, v10, v11, v12* i *v13*. Wybrano jedynie podzbiór widoków na środku sceny aby zmniejszyć złożoność obliczeniową. Następnie, na ich podstawie syntezowano widok wirtualny w pozycji widoku *v9*. Dzięki takiemu podejściu, możliwa była końcowa ocena jakości renderowanej treści z użyciem dwóch obiektywnych metryk jakości dedykowanych do oceny wizji wszechogarniającej: IV-PSNR [7] i SSIV [22]. W eksperymencie sprawdzono wpływ:

- przeprowadzenia odszumiania sekwencji (z użyciem algorytmu NL-means [3]),
- korekcji barwnej (algorytmem PCR [6]),
- poprawności wyznaczenia parametrów zewnętrznych (dwa testy: dokładność parametrów 1,5 i 2 razy mniejsza, niż w przypadku optymalnych, ostatecznych parametrów wyznaczonych algorytmem ECPC [23]),
- dwukrotnego zmniejszenia liczby kamer (użycie wyłącznie widoków *v6, v8, v10* i *v12*),
- niepoprawnej synchronizacji kamer [dwa testy: jedna kamera rejestrowała scenę z opóźnieniem jednej (40 ms) lub dwóch ramek (80 ms)],
- niepoprawnej korekcji zniekształceń soczewkowych (użycie parametrów zniekształcenia większych o 1 % od rzeczywistych).

Wyniki eksperymentu zaprezentowano na rysunku 5. Jak pokazano, każdy z opisanych etapów przetwarzania sekwencji wielowidokowej ma istotny wpływ na ostateczną jakość syntezowanych widoków. Dokładność synchronizacji i kalibracji systemu wielokamerowego wraz

z przeprowadzeniem niezbędnej korekcji zarejestrowanego materiału jest kluczowa z punktu widzenia ostatecznego użytkownika systemu swobodnej nawigacji.



Rys. 5. Jakość syntezowanego widoku wirtualnego dla różnych konfiguracji procesu przetwarzania sekwencji wielowidokowej.

4. PODSUMOWANIE

W niniejszym artykule zaprezentowano przegląd metod wykorzystywanych w procesie tworzenia naturalnych sekwencji wielowidokowych na potrzeby wizji wszechogarniającej. Tak przygotowana sekwencja wizyjna jest gotowa do przedstawienia końcowemu użytkownikowi na ekranie telewizora czy też z wykorzystaniem nagłownych okularów do rzeczywistości wirtualnej [2]. Wyniki przeprowadzonych eksperymentów pokazują wpływ poszczególnych kroków procesu akwizycji wizji wszechogarniającej na końcową jakość.

Proces akwizycji i tworzenia testowej sekwencji wielowidokowej z wykorzystaniem przytoczonych metod został przedstawiony na przykładzie sekwencji Choreo, która została zgłoszona przez Autorów do bazy sekwencji testowych standardu ISO/IEC MPEG Immersive Video, i która to jest obecnie wykorzystywana w testach MPEG INVR.

PODZIĘKOWANIA

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

LITERATURA

[1] H. Boukamcha. i in. 2017. "Robust auto calibration technique for stereo camera". *ICEMIS 2017*.

[2] Boyce J. M. i in. 2021. "MPEG Immersive Video Coding Standard". *Proc. of the IEEE* 109: 1521-1536

[3] Buades A. 2005. "A non-local algorithm for image denoising". *CVPR 2005*: 60–65.

[4] Cui H. i in. 2023. "MCSfM: Multi-Camera Based Incremental Structure-From-Motion". *IEEE Transactions on Image Processing*, 32: 6441 - 6456.

[5] Domanski M. i in. 2016. „New results in free-viewpoint television systems for horizontal virtual navigation". *ICME 2016.*

[6] Dziembowski A. i in. 2021. „Color Correction for Immersive Video Applications". *IEEE Access* 9: 75626-75640.

[7] Dziembowski A. i in. 2022. „IV-PSNR–the objective quality metric for immersive video applications". *IEEE T. Circ. & Sys. Vid. Tech.* 32 (1): 7575-7591.

[8] Ghorbanzadeh D. i in. 2024. "Enhancing destination image through virtual reality technology: the role of tourists' immersive experience". *Curr. Psych:* 1-13.

[9] Hinds T. A. i in. 2023. "Immersive Media and the Metaverse". *IEEE Comm. Magazine* 61: 48-54.

[10] ISO/IEC SC29/WG04, 2024, "Manual of the Extrinsic Camera Parameters Calibration framework". *146 MPEG meeting,* Rennes, Francja.

[11] Jianzhu H. i in. 2023. "A Review and Comparative Study of Close-Range Geometric Camera Calibration Tools". arXiv:2306.09014.

[12] Kang Y. i in. 2010. "High-quality multi-view depth generation using multiple color and depth cameras". *ICME 2010*, pp. 1405–1410.

[13] Klóska D. i in. 2024. „Proposal of new natural content – Choreo". *MPEG 146,* Rennes, Francja.

[14] Kolmogorov V. i in. 2004 "What Energy Functions Can Be Minimized via Graph Cuts?". *IEEE Tr. on Pattern Analysis and Machine Int.* 26.2: 147–159.

[15] Liu R. i in. 2009. "Stereo Cameras Self-Calibration Based on SIFT". *Int. Conf. on Measuring Tech. and Mechatronics Automation*, 1: 352-355.

[16] Meng X. i in. 2018. "Dense RGB-D SLAM with Multiple Cameras". *Sensors*, 18 (7): 2118.

[17] Mieloch D. i in. 2020. „Depth map estimation for free-viewpoint television and virtual navigation". *IEEE Access* 8: 5760–5776.

[18] Mueller K. i in. 2011. „3-D Video Representation Using Depth Maps". *Proc. of the IEEE* 99: 643–656.

[19] Pengwei Z. i in. 2024. "Meta-Calib: A generic, robust and accurate camera calibration framework with ArUco-encoded meta-board". *ISPRS Journal of Photogrammetry & Remote Sensing* 212: 357-380.

[20] Stankiewicz O. i in. 2018. "A Free-Viewpoint Television System for Horizontal Virtual Navigation". *IEEE Transactions on Multimedia* 20 (8): 2182-2195.

[21] Stankiewicz O. i in. 2018. "Multiview video: Acquisition, processing, compression and virtual view rendering". *Acad. Press Lib. in Sig. Proc.* 6: 3-74.

[22] Stankowski J. i in. 2024. „Miara podobieństwa strukturalnego dla wizji wszechogarniającej". *KRiT 2024.*

[23] Szydełko B. i in. 2024. „ECPC – versatile multi-camera system calibration framework for immersive video applications". *SoftwareX,* 26 (101669).

[24] Wei Z. i in. 2023. „Research on the Current Situation and Future Development Trend of Immersive VR in the Field of Education". *Sustainability,* 15, 7531.

[25] Wien M. i in. 2019. "Standardization Status of Immersive Video Coding". *IEEE J on Emerging & Selected Topics in Circuits & Systems* 9 (1): 5-17.

[26] Xiang S. i in. 2013. „A gradient-based approach for interference cancelation in systems with multiple Kinect cameras". *ISCAS 2013.*

[27] Yao Y. i in. 2018. „MVSNet: Depth Inference for Unstructured Multi-view Stereo". *ArXiv*1804.02505.

[28] Zhang Z. 2000. "A Flexible New Technique for Camera Calibration*". IEEE Transactions on Pattern Analysis & Machine Intelligence.* 22 (11): 1330 - 1334.

[29] Zizhuang W. i in. 2021. „AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network". *ArXiv,* 2108.03824.