Raport Badawczy

Research Report

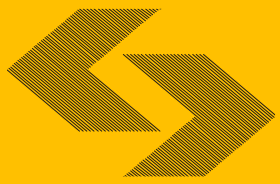
RB/61/2010

Wielo-kadrowa analiza obrazów w zagadnieniach rozszerzonej rzeczywistości

K. Koniarski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



WIELO-KADROWA ANALIZA OBRAZÓW W ZAGADNIENIACH ROZSZERZONEJ RZECZYWISTOŚCI

Konrad Koniarski
Studia Doktoranckie IBS PAN

The paper deals with the recognition of the same or similar objects in the different movie sequences. The recognition algorithm based on epipolar geometry constraints is presented. This algorithm consists from three main functional blocks. The first block selects characteristic points of the frame using the corner detection methods. The second block employs RANSAC method to eliminate outliners among the selected points that cannot be matched between frames. The last block determines the location of real image points using epipolar geometry as well as the minimization of a distance function. Levenberg – Marquad algorithm is used as optimization method. Moreover the application of the described algorithm in Augmented Reality problems is discussed.

Key words: *Augmented reality, epipolar geometry, optimization methods, Levenberg-Marquad algorithm*

Wstęp

Kamery cyfrowe są powszechnie używane do rejestrowania obrazów. W tym celu wykorzystuje się zarówno specjalistyczne kamery wideo lub monitoringu przemysłowego jak i kamery używane w telefonach komórkowych. Urządzenia te jednak jedynie tworzą sygnał cyfrowy reprezentujący obraz i pozwalają na jego utrwalenie. Wszelkie interpretacje tego, co przedstawia obraz pozostawione są nadal wyobraźni obserwatora. Powstało już wiele metod bazujących głównie na segmentacji i metodzie aktywnego konturu [4, 6, 7, 9, 12, 13], które wspomagają proces analizy obrazu. Metody te mają na celu selekcję, wyróżnienie i zaprezentowanie istotnych elementów obrazu osobie go interpretującej. Obecnie obszarem intensywnych badań [2, 3, 5, 14] są również metody przetwarzania sekwencji obrazów cyfrowych, gdyż wiele istotnych informacji może nie być zauważalnych na pojedynczym kadrze.

Celem artykułu jest omówienie algorytmu, który wyodrębni obiekty sceny przedstawione w sekwencji filmowej oraz pozwoli na wyznaczenie ich wzajemnego położenia. Algorytm ten nie tylko analizuje z osobna każdy kadr, ale

też wychwytuje cenne informacje, które znajdują się „pomiędzy" kadrami. Zasadniczymi kryteriami użyteczności przedstawionego algorytmu są szybkość działania (działanie w czasie rzeczywistym) oraz jakość uzyskanych rezultatów w postaci prawidłowo wyodrębnionych obiektów.

1. Podstawowe pojęcia rozszerzonej rzeczywistości

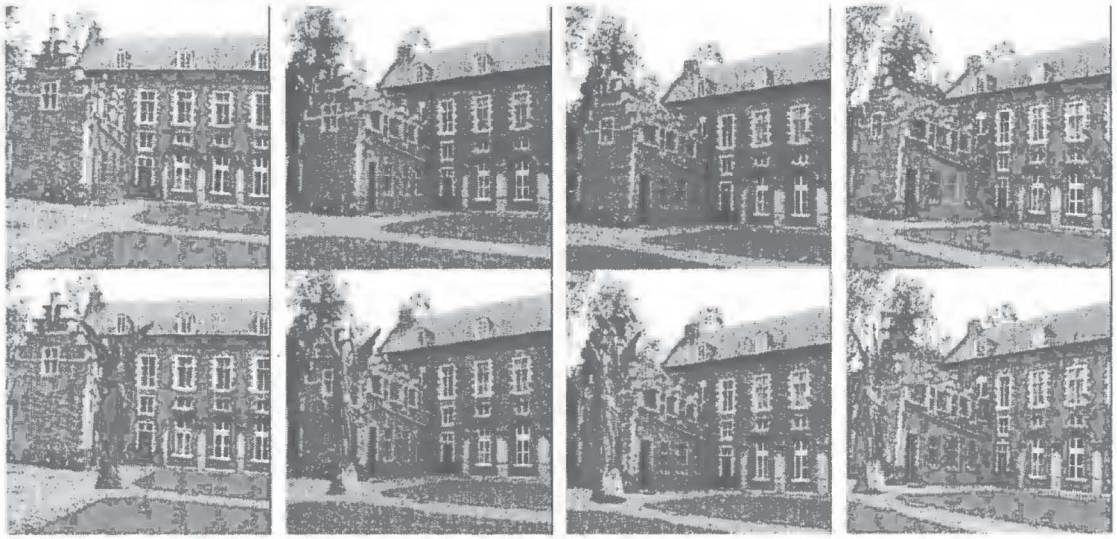
Rozszerzona rzeczywistość (ang. Augmented Reality, AR) to zbiór technik wykorzystywanych w informatyce do uzupełniana przekazu multimedialnego o dodatkowe treści. Rozszerzanie rzeczywistości może odbywać się poprzez modyfikację obrazu lub dźwięku. W definicji podanej przez Ronalda Azuma [1] jeszcze w latach 90. dwudziestego wieku wyróżnia się trzy podstawowe elementy, które każdy system AR musi zawierać. Po pierwsze AR łączy obraz rzeczywisty i wygenerowany komputerowo tak, że oba przenikają się wzajemnie. Po drugie system działa w czasie rzeczywistym. Po trzecie system działa w otoczeniu trójwymiarowym umożliwiając swobodne przemieszczanie się kamery oraz obiektów filmowanej sceny.

AR w dziedzinie obrazu odnosi się do zbioru technik, które pozwalają na analizę i modyfikację obrazu poprzez dodanie elementów wirtualnych bądź usunięcie elementów istniejących. Najczęściej terminu AR używa się w kontekście kompozycji obiektów wirtualnych w sekwencji wideo przedstawiającej obiekty rzeczywiste [1]. Takie zastosowanie niesie za sobą dwa ograniczenia, które muszą być spełnione, aby powstały obraz był rzeczywisty dla odbiorcy. Pierwszym ograniczeniem jest takie umiejscowienie obiektu wirtualnego w scenie obrazu rzeczywistego, aby nie naruszał on zasad kompozycji np. zasłaniając obiekty bliższe kamerze niż on sam. Drugim ograniczeniem jest rozmieszczenie pozostałych obiektów wirtualnych względem istniejących obiektów rzeczywistych z zachowaniem zasad kompozycji.

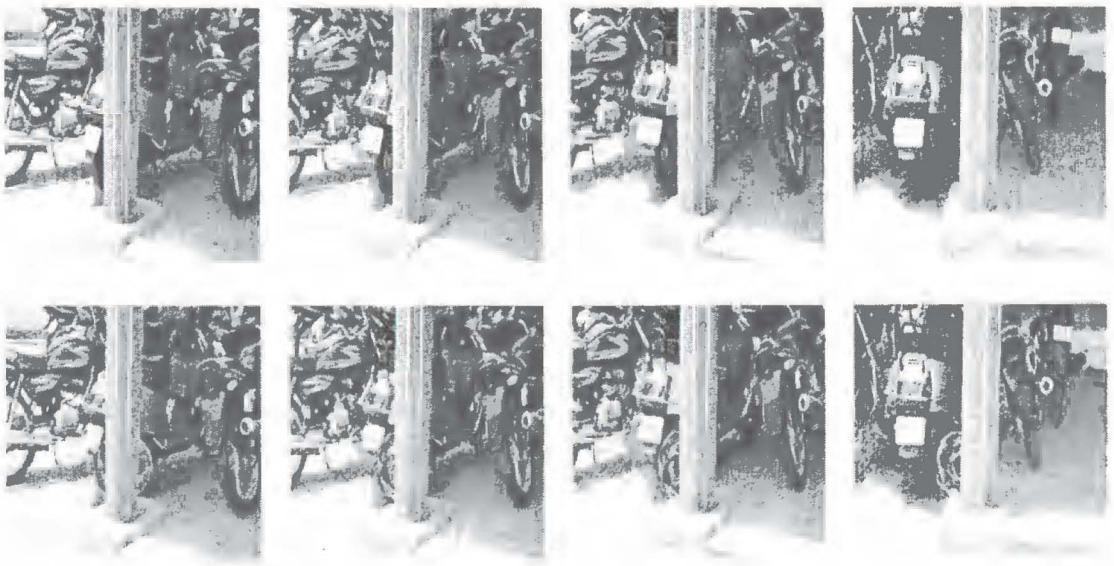
Rysunek 1. przedstawia przykład użycia technik AR. W górnym wierszu przestawione są wybrane kadry z sekwencji filmowej. Natomiast w dolnym wierszu przedstawione są te same kadry, do których dodano obiekt sztuczny stworzony przy pomocy grafiki 3D. Przykład dobrze obrazuje właściwie usytuowanie obiektu wirtualnego przy ruchu kamery. Jak widać zmieniające się położenie kamery powoduje zmianę kąta, pod którym wirtualny obiekt jest oglądany.

Rysunek 2. ilustruje zastosowanie AR do wstawienia obiektu wirtualnego, tak, że jest on w części przysłaniany przez obiekt pierwszego planu. Wykorzystanie informacji o głębi obrazu z kolejnych kadrów pozwala na określenie,

które obiekty znajdują się bliżej kamery. Dzięki czemu możliwe staje się wstawienie obiektów wirtualnych w taki sposób, aby pierwszoplanowe elementy sceny przykrywały obiekt wirtualny.



Rysunek 1. AR wstawienie obiektu wirtualnego w sekwencji kadrów [5]



Rysunek 2. AR wstawienie obiektu pomiędzy obiekty sceny [5]

W dziedzinie obrazu AR możemy podzielić na techniki wykorzystujące znacznik (ang. marker) [6] i te, które nie potrzebują znacznika (ang. markerless) [2, 9]. Metody niewykorzystujące znacznika powstały znacznie wcześniej, ze względu na mniejszą złożoność algorytmów wykorzystywanych w tej technice. Znacznik to niesymetryczny płaski obraz wprowadzony do filmowanej sceny w celu utworzenia macierzy kamery P (ang. camera matrix).

$$X_{kadru} = PX_{sceny}, \tag{1}$$

gdzie X_{sceny} to położenie punktu w przestrzeni $\mathbb{R}^3$, w której jest filmowana scena. Z kolei X_{kadru} to położenie punktu na kadrze reprezentowanym przez przestrzeń $\mathbb{R}^2$.

Macierz kamery definiuje sprzętowe własności kamery związane ze zniekształceniami powstałymi w wyniku otrzymywania obrazu. Tym samym macierz kamery opisuje operacje translacji i rotacji, jakim należy poddać sztuczny obiekt przed włączeniem go do obrazu rzeczywistego. Niestety wprowadzenie znacznika wymaga ingerencji w zbiór filmowanych obiektów, co nie zawsze jest możliwe. Kolejną wadą wykorzystania metod bazujących na rozpoznaniu znacznika jest to, że znacznik może być zasłonięty przez inne elementy sceny, co uniemożliwia dodanie obiektów wirtualnych.

Metody bez znacznika tworzą macierz kamery z informacji zawartych w samym obrazie. Analiza kilku kadrów z filmowanej sekwencji pozwala na wyznaczenie wzajemnego położenia obiektów w scenie. Dzięki temu możliwe jest nałożenie obiektów wirtualnych na rzeczywiste z uwzględnieniem wszystkich zasad kompozycji tworząc w odbiorze obserwatora obraz naturalny.

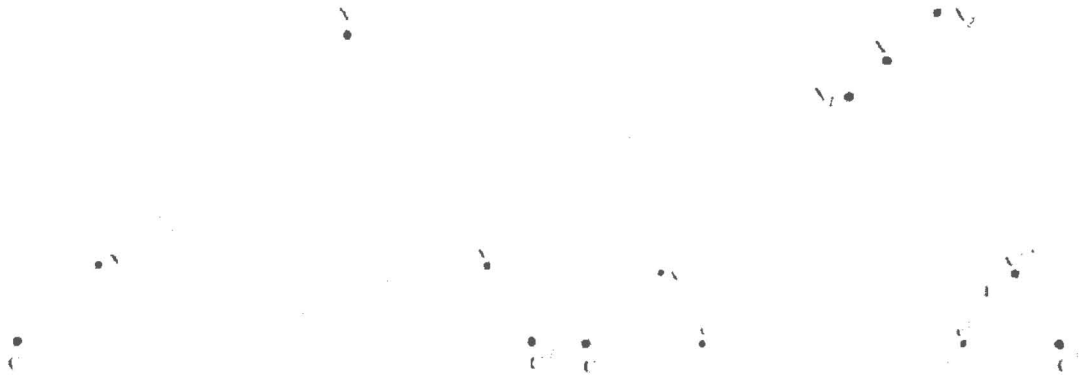
2. Geometria epi-biegunowa

Przy analizie wielu widoków tej samej sceny nie ma znaczenia czy jest to obraz uzyskany przez dwie kamery czy jedną kamerę z wielu ujęć. Oba obrazy powiązane są zależnościami, które opisuje geometria epi-biegunowa (ang. epipolar geometry) [9]. Dla wygody czytelnika poniżej zostaną przypomniane podstawowe pojęcia i zależności tej geometrii w oparciu o pracę [9].

Załóżmy, że obserwujemy wybrany punkt X w przestrzeni trójwymiarowej (patrz Rysunek 3. Płaszczyzna epi-biegunowa). Jego reprezentacją na płaszczyznach obrazów są punkty x i x', odpowiednio, dla pierwszej i drugiej kamery. Punkt X i jego reprezentacje x i x' na kadrach obrazu są powiązane równaniami:

$$x = PX \text{ oraz } x' = P'X', \tag{2}$$

gdzie P i P' to macierze kamer [9, 8]. Punkty reprezentujące środki kamer C i C', punkt X oraz obrazy punktu X na płaszczyznach kadrów, odpowiednio, x oraz x' leżą na jednej płaszczyźnie zaznaczonej na rysunku ciemnym kolorem, zwanej płaszczyzną epi-biegunową (ang. epipolar plane).



Rysunek 3. Płaszczyzna epi-biegunowa [9]

Załóżmy, że przesuwamy punkt X wzdłuż odcinka XC. Położenie punku x na płaszczyźnie kadru nie zmieni się. Natomiast kolejny rzut punktu X, punkt x', przesunie się wzdłuż prostej l'. Jak widać na postawie samego rzutu punku X do punku x nie jest możliwe ustalenie dokładnego położenia punku X w przestrzeni. Każdy punku leżący na półprostej o początku w punkcie C i przechodzącej przez punkt x (patrz Rys 2. Punkty X_1, X_2, X) będzie rzutowany do punktu x. Zbiór kolejnych położeń punktów x' tworzy jeden odcinek zwany linią epipolarną (ang. epipolar line). Punkty e i e' leżące na przecięciu linii epipolarnych i środków kamer C i C' nazywamy epi-biegunami (ang. epi-polami).

W geometrii epi-polarnej do określenia położenia punktów x oraz linii l jest wykorzystywana macierz fundamentalna F (ang. fundamental matrix). Wyznaczenie tej macierzy możemy rozpocząć od wyznaczenie zależności pomiędzy danym punktem x a jego rzutem prostokątnym na linię epi-biegunową l' na drugim kadrze. W tym celu załóżmy, że obserwowany punkt X leży na płaszczyźnie π. Jeśli punkt X leży na półosi o początku x, jego rzut x' musi leżeć na linii epi-biegunowej l'. Oznaczając przez H_π transformacje punktu x na x' możemy wtedy zapisać:

$$x' = H_\pi x \text{ oraz } l' = [e']_\times x' = [e']_\times H_\pi x = Fx, \tag{3}$$

gdzie

$$F = [e']_\times H_\pi, \tag{4}$$

jest macierzą fundamentalną. Zauważmy, że drugie równanie w (3) oznacza, że dany punkt x', linia epibiegunowa l' przechodząca przez ten punkt obrazu oraz epi-biegun e' są od siebie zależne.

Geometrycznie, macierz F przedstawia rzutowanie z płaszczyzny jednego kadru na linię epi-biegunową na kadrze drugim przechodzącą przez epi-biegun e'. Analogicznie możemy rzutować punkty z drugiego kadru na pierwszy. Wówczas

$$F = [e]_X H_\pi \text{ oraz } x' = H_\pi x \ , l = [e]_X x' = [e]_X H_\pi x' = F x', \quad (5)$$

a punkty x i x' powiązane są przy pomocy macierzy F. Dla każdej pary punktów x i x' spełniony jest warunek ortogonalności [9]

$$x'^T F x = 0. \quad (6)$$

Linia epi-polarna $l' = Fx$ zawiera epibiegun e'. Zatem

$$e'^T (Fx) = (e'^T F)x = 0, \quad (7)$$

gdzie e^T oznacza transpozycję wektora e. Położenie epi-biegunów możemy wyznaczyć z zależności:

$$e'^T F = 0 \wedge Fe = 0. \quad (8)$$

Wykorzystując własności geometrii epi-polarnej możliwe jest uzyskanie dokładnych informacji o położeniu obiektów względem kamer oraz względem siebie. Pozwala to na dowolne dalsze zastosowanie pozyskanych w ten sposób danych np. w AR czy budowie mapy głębi sceny.

Niestety obrazy uzyskane z kamer cyfrowych zawierają zazwyczaj szum Gaussowski (ang. Gaussian noise) [6, 9]. Ponieważ szum ten powoduje, że półosie poprowadzone ze środków kamer C i C', odpowiednio, przez punkty x i x' nie przetną się dokładnie w punkcie X zatem bezpośrednie wyznaczenie macierzy fundamentalnej powyższą metodą nie jest w ogólnym przypadku możliwie. Natomiast wiemy, że odległość, w jakich półosie przechodzą obok siebie nie jest duża. Pozwala to na określenie pewnego sąsiedztwa punktu X, przez które z pewnością przechodzą obie półosie.

3. Algorytm wyszukiwania podobnych obiektów

Algorytm wyszukiwania podobnych obiektów w różnych kadrach sekwencji filmowej składa się z kilku bloków funkcjonalnych. Każdy taki blok

można implementować różnymi metodami. Schemat blokowy tego algorytmu przedstawia Diagram 1.

Pierwszym krokiem algorytmu jest wyodrębnienie charakterystycznych punktów obrazu. Należy tu zwrócić uwagę na to, aby punkty były tak samo rozpoznawane niezależnie od perspektywy, z której są obserwowane. W tym celu stosuje się metody wykrywania kątów przedstawionych obiektów (ang. corner detection methods) takie jak operator Moravec'a, operator Harrisa/Pleesey'a czy operator Haralick'a [4, 7]. Zasada działania tych operatorów jest zbliżona. Wszystkie wykorzystują metodę ruchomego okna i autokorelacji zadanej maski wykrywającej z obrazem.

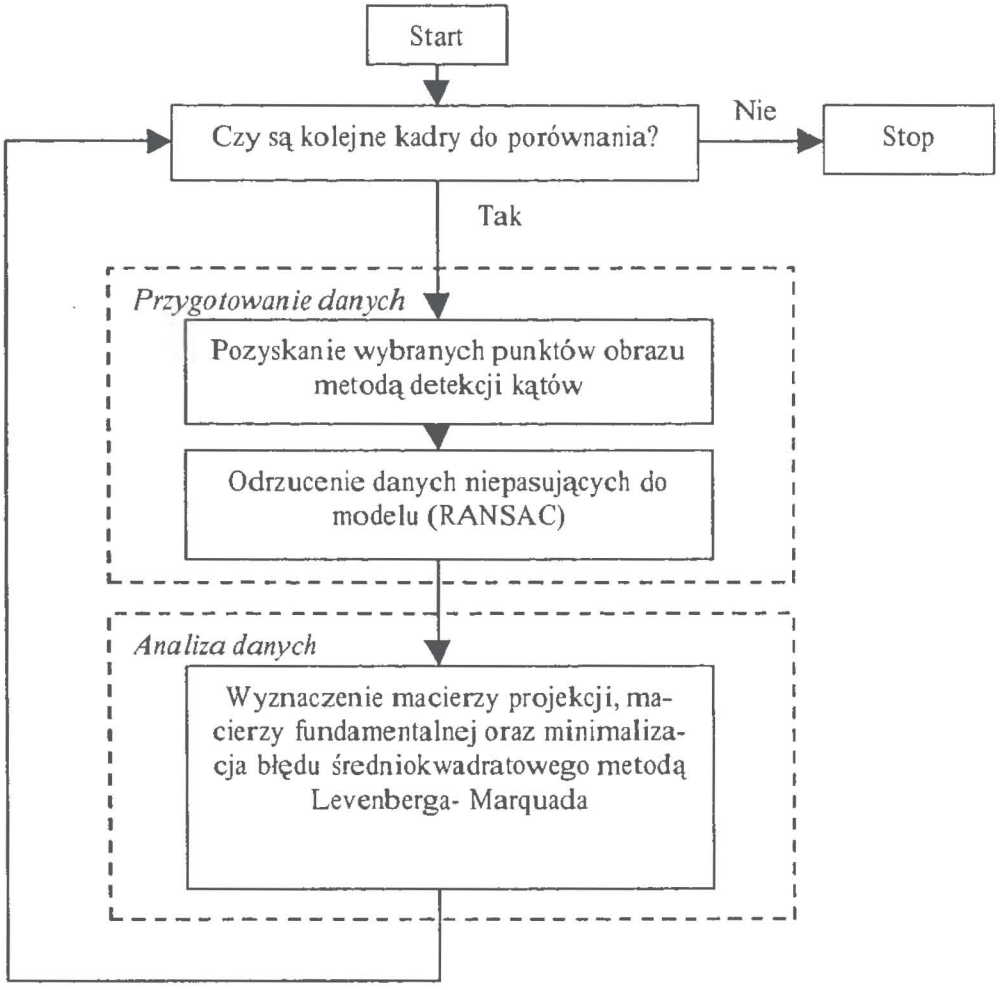


Diagram 1. Schemat blokowy algorytmu

W następnym kroku algorytmu następuje analiza danych uzyskanych z obr zu i odrzucenie danych nadmiarowych. Implementacja tego bloku algoryt' a z powodzeniem może być wykonana przy użyciu metody RANSAC (Random Sample Consensus) [9]. Metoda RANSAC to metoda iteracyjna eli-

minująca losowo, przy założonym progu tolerancji, wartości parametrów niepasujących do zakładanego modelu obrazu.

Ponieważ metoda RANSAC korzysta z losowości przy odrzucaniu paramentów, zatem jej czas działania może być długi. Z drugiej strony możliwe jest też parametryzowanie czułości detekcji interesujących punktów. Nie rozwiązuje to jednak problemu odrzucania istotnych punktów (ang. inliers), które reprezentują zasadnicze informacje o rozmieszczeniu obiektów na scenie.

W kolejnym bloku algorytmu następuje analiza zebranych danych poprzez dopasowanie do siebie punktów z różnych kadrów przedstawiających ten sam obiekt rzeczywisty. W tym celu wyznaczana jest macierz fundamentalna F. Do wyznaczenia macierzy potrzebne jest przynajmniej 8 punktów wybranych z obrazu [9], które sobie odpowiadają $\{x_i \longmapsto x'_i\}$. W tym celu poddajemy punkty x_i i x'_i transformacjom normalizującym T oraz T', które uwzględniają skalowanie i translacje. Transformacje T oraz T' są znane. Otrzymujemy

$$\hat{x}_i = Tx_i \text{ oraz } \hat{x}'_i = T'x'_i, \tag{9}$$

gdzie $\hat{x}_i$ i $\hat{x}$ są punktami po normalizacji. Następnie znajdujemy macierz $\hat{F}$, która łączy punkty $\{\hat{x}_i \longmapsto \hat{x}'_i\}$. W tym celu wprowadźmy oznaczenia $x = (x, y, 1)^T, x' = (x', y', 1)^T$. Korzystając z równania (6) możemy zbudować układ równań z niewiadomą F. Jego rozwiązanie będzie wyglądało wtedy następująco

$$x'xf_{11} + x'yf_{12} + x'f_{13} + y'xf_{21} + y'yf_{22} + y'f_{23} + xf_{31} + yf_{32} + f_{33} = 0. \tag{10}$$

Przez f oznaczony jest wektor, z którego budujemy macierz fundamentalną wstawiając najpierw wiersze, a później kolumny. Wówczas powyższe równanie możemy zapisać w postaci

$$(x'x, x'y, x', y'x, y'y, y', x, y, 1)f = Af = 0. \tag{11}$$

Określamy $\hat{F}'$ na postawie wektora odpowiadającego najmniejszej wartości $\hat{A}$, gdzie $\hat{A}$ jest tworzone z punktów $\{\hat{x}_i \longmapsto \hat{x}'_i\}$. Następnie, korzystając z metody SVD, zastępujemy $\hat{F}$ przez $\hat{F}'$. Macierz $\hat{F}$ jest macierzą najbliższą macierzy $\hat{F}'$ w metryce Frobeniusa [9]. Ostatnim etapem algorytmu wyznaczania macierzy fundamentalnej F jest denormalizacja $F = T'^T \hat{F}' T$. W jej wyniku otrzymujemy macierz F łączącą punkty $\{x_i \longmapsto x'_i\}$.

Kolejny krok algorytmu eliminuje niedokładność pomiaru. Zakładamy, że pomiar punktu obrazu podlega szumowi o rozkładzie Gaussowskim, to znaczy jest niedokładny. Celem tego bloku jest wyznaczenie najlepszego dopasowania pomiędzy punktami x i X, oraz x' i X.

W tym celu aby wyznaczyć punkty $\hat{x}_i$ i $\hat{x}'_i$ minimalizujemy poniższą funkcję odległości,

$$\sum_i d(x_i,\hat{x}_i)^2 + d(x'_i,\hat{x}'_i)^2, \tag{12}$$

względem zmiennych $\hat{x}_i$ i $\hat{x}'_i$, które spełniają ograniczenie

$$\hat{x}'^{T}_i F \hat{x}_i = 0, \tag{13}$$

gdzie x_i i x'_i są punktami z pomiarów a d oznacza odległość punktów x_i i $\hat{x}_i$ w metryce Euklidesowej zaś $i = 1,\ldots, N$, gdzie N jest zadaną liczbą naturalną.

Do minimalizacji wartości funkcji (12) jest wykorzystywany algorytm Levenberga-Marquardta [11]. Ponieważ algorytm ten łączy cechy metody największego spadku i metody Gaussa-Newtona dlatego zapewnia on szybką zbieżność.

Każdy kolejny znaczący kadr obrazu (ang. key frame) jest poddawany analizie. Tak uzyskane dane o wzajemnych relacjach poszczególnych punktów pomiędzy kadrami są podstawą do dalszej analizy lub zastosowania w AR. Pozwala to na zebranie przestrzennych zależności filmowanej sceny.

4. Zastosowanie

Omówiony algorytm pełni zasadniczą rolę w dalszej analizie sceny obrazu lub w zastosowaniu do AR. Szczególnie istotnym zastosowaniem wydaje się wyznaczanie głębi obrazu na podstawie sekwencji filmowej. Podobne zadanie zostało sformułowane w pracy [10] gdzie wykorzystana metoda opierała się na segmentacji obrazu stereoskopowego. Cechą odróżniającą zaproponowany w tej pracy algorytm od algorytmu przedstawionego w pracy [10] jest spełnienie wymogów stawianych w metodach AR tj., działanie w czasie rzeczywistym. Jednocześnie wykorzystanie metody RANSAC i algorytmu Levenberga-Marquada poprawia wydajność zaproponowanego algorytmu. Wydajność oraz uniwersalność proponowanego algorytmu pozwalają na jego zastosowanie w urządzeniach nieposiadających dużej mocy obliczeniowej. Przykładem takich urządzeń są telefony komórkowe typu smartphone wyposażone w kamerę.

Podsumowanie

Ilość zastosowań metod AR bez użycia znacznika jest bardzo duża, o czym świadczy ciągle rosnące zainteresowanie różnymi modyfikacjami metody [14].

Wykorzystanie geometrii epi-polarnej w omówionym algorytmie pozwala na swobodne przemieszczanie zarówno obiektów jak i samej kamery. Metoda nie posiada znacznych ograniczeń zastosowania tak jak metody AR wykorzystujące znacznik, ani metoda przedstawiona w [10], gdzie wykorzystywana było specyficzna pozycja kamer. Zastosowanie szybkich metod optymalizacji nieliniowej jest daje możliwość użycia przedstawionej metody do rozwiązywania problemów w czasie rzeczywistym.

Podobna metoda została przedstawiona w artykule [5]. Obie metody wykorzystują podobne metody detekcji punktów oraz metodę RANSAC. Po wykonaniu obliczeń, stosowne będzie porównanie zaprezentowanej metody z metodą wykorzystującą segmentację.

Zasadniczym celem przyszłych badań dotyczących zaprezentowanej metody są dwa pierwsze bloki funkcjonalne opisanego algorytmu odpowiadające za ekstrakcję cech obrazu oraz wstępną selekcję dopasowania do modelu. Celem dalszych prac będzie określenie, jak należy parametryzować algorytm w tych blokach, aby zwiększać wydajność całej zaproponowanej procedury. Ponadto rozważana jest zmiana metryki funkcji odległości pomiędzy punktami rzeczywistymi a wyliczonymi, co może mogą mieć wpływ, na jakość wyników algorytmu w ostatnim bloku funkcjonalnym zaprezentowanej metody.

Literatura

[1] Azuma, R. T. (1997): A Survey of Augmented Reality. *Teleoperators and Virtual Environments*, 6(4), 355-385.

[2] Bilton, N. (2010): *G.M. Tinkers With Augmented Reality System for Cars*, The New York Times, March 17, (link: http://bits.blogs.nytimes.com / 2010 / 03 / 17/gm - tinkers - with –augmented – reality – system – for -cars/?scp=1&sq= augmented%20 reality&st=cse).

[3] Bleser, G. Becker, M. Stricker, D. (2007): Real – time vision – based tracking and reconstruction. *J. Real – Time Image Proc.*, 2, 161-175.

[4] Bradski, G. Kachler, A. (2008): *Learning OpenCV*, O'Really Media Inc. CA.

[5] Chari, V. Singh, J. M. Narayanan, P. J. (2004): *Augmented Reality using Over-Segmentation*. Preprint.

[6] Davies, E. R. (2005): *Machine Vision, Theory, Algorithms, Practicalities*, Elsver Inc. San Francisco.

[7] Fischer, R. B. (2007): *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*, Preprint, (link: http://homepages.inf.ed.ac.uk/rbf/Cvonline/feature.htm).

[8] Hartley, R.I. Sturm, P. (1997): Triangulation. *Computer Vision and Image Understanding*, 68(2), 146-157.

[9] Hartley, R. I. Zisserman, A. (2003): *Multiple View Geometry In Computer Vision*. Cambridge University Press, Cambridge, England.

[10] Koniarski, K. (2009): *Stereoskopowa segmentacja obrazu*. W: Analiza systemowa w finansach i zarządzaniu. Wybrane Problemy, Tom 11, red. J. Hołubiec. Instytut Badań Systemowych PAN, Warszawa, 148-152.

[11] Nocedal, J. Wright, S. J. (2000): *Numerical Optimization*, Springer, New York, 259-269.

[12] Russ, J. C, (2007): *The Image Processing Handbook*, 5th edition, Taylor & Francis `Group, Raleigh.

[13] Seul, M. O'Gorman, L. Sammon M. J. (2000): *Practical Algorithms for Image Analysis*, Cambridge, Cambridge University Press.

[14] Schonfeld, E. (2010): *Augmented Reality Vs. Virtual Reality: Which One Is More Real?*, (link: http://techcrunch.com/2010/01/06/ augmented-reality-vs-virtual-reality/