dr hab. inż. Leszek J. Chmielewski, prof. SGGW
Instytut Informatyki Technicznej
Szkoła Główna Gospodarstwa Wiejskiego
ul. Nowoursynowska 159, 02-776 Warszawa

7 września 2022 r.

Recenzja rozprawy doktorskiej

autor rozprawy:

mgr inż. Marek Kulbacki

Wydział Informatyki, Polsko-Japońska Akademia Technik Komputerowych

tytuł rozprawy:

Learning Scene Dependent Models of Human Actions and Behavior from Surveillance Video

Recenzja została przygotowana w odpowiedzi na powołanie, przez Radę Naukową Dyscypliny Informatyka Polsko-Japońskiej Akademii Technik Komputerowych, na recenzenta rozprawy doktorskiej Pana mgra inż. Marka Kulbackiego. Promotorem rozprawy jest prof. dr hab. inż. Konrad Wojciechowski, a promotorem pomocniczym dr Jakub Segen. Przewód jest prowadzony w dyscyplinie *informatyka techniczna i telekomunikacja.*

1 Zawartość rozprawy

1.1 Charakterystyka ogólna

W recenzowanej dysertacji zaproponowano całościowy zbiór metod oraz oprogramowanie na wysokim poziomie zaawansowania technologicznego do rozpoznawania akcji człowieka na podstawie rzeczywistych danych wideo pochodzących z bezpośredniego monitoringu lub innych źródeł. Możliwe jest rozpoznawanie wielu jednocześnie zachodzących indywidualnych akcji wielu osób, ich lokalizację w czasie i w przestrzeni oraz klasyfikację.

Rozwiązano problem kalibracji w czasie rzeczywistym, co pozwala na detekcję obiektów w rzeczywistych współrzędnych świata w sposób ciągły, przy jednoczesnych zmianach położenia (kątów) i ogniskowej kamer, przy czym jednocześnie zachodzi korekcja nieliniowych zniekształceń kamer. Korzysta się w tym celu z informacji o położeniach punktów leżących na płaszczyźnie podłoża.

Opracowano system opisu sceny wraz z grupami trajektorii postaci ludzkich i innych ruchomych obiektów. Ponieważ nie korzysta się ze szkieletowego modelu postaci ludzkiej, to opis jest prosty i nie ma ograniczeń na typ poruszającego się obiektu. W zaproponowanej reprezentacji korzysta się z symboli, które sa łączone w sekwencje mające parametry geometryczne i czasowe. Dzięki temu można wydzielić ruchy poszczególnych osób w sekwencji wideo.

Na podstawie takiej reprezentacji można rozpoznawać akcje osób i grup osób w czasie zbliżonym do rzeczywistego. Użyto do tego celu lokalnych reprezentacji czasoprzestrzennych oraz zoptymalizowanego podejścia *Bag of Visual Words.*

Stworzone algorytmy przetestowano na danych rzeczywistych z publicznie dostępnych zbiorów danych wideo: Hollywood2, UCF101, HMDB51, oraz własnych baz VMASS i VMASS2 z miejskich sieci kamer monitoringu.

Opracowane metody wykorzystano w autorskim systemie monitoringu SAVA. Jest to system klasy przemysłowej. Został on przetestowany z wynikiem pozytywnym na wyżej wymienionych bazach. Osiągnięto czasy zbliżone do czasu rzeczywistego w systemie monitoringu pracującym w rzeczywistych warunkach. Między innymi, osiągnięto bardzo dobre wyniki rozpoznawania typów akcji, na przykład, osiągnięto dokładność wykrywania akcji potencjalnie niebezpiecznych ponad 70%.

1.2 Omówienie części pracy

Rozprawa jest napisana w języku angielskim. Składa się z 12 rozdziałów podzielonych na dwie części, spisu akronimów, bibliografii oraz załącznika, który jest jednocześnie trzecią częścią pracy i zawiera trzy duże rozdziały podzielone na siedem rozdziałów.

Pierwsze trzy rozdziały to skróty pracy – w języku angielskim i polskim, oraz podziękowania.

W rozdziale czwartym, który jest wstępem, uzasadniono przeprowadzone badania, oraz przedstawiono zakres pracy. Nie postawiono tezy, natomiast zdefiniowano trzy cele pracy, których sformułowania można uznać za równoważne tezom. Podano także trzy główne wymagania, które musi spełnić oprogramowanie, w którym zrealizowano te cele. Omówiono własne publikacje Doktoranta, które były podstawą dla rozprawy. Omówiono także zawartość poszczególnych części pracy.

Rozdział piąty zawiera omówienie stanu wiedzy, z podziałem na inteligentne metody analizy sekwencji wideo oraz rozpoznawania akcji człowieka, z głębszym podziałem. Opisy zawarte w tym rozdziale są bardzo szeroko zakrojone, a jednocześnie dogłębne. Zastosowano się do standardu *Intelligent Video Analytics* (IVA), który jest powszechnie stosowany w rozwiązywaniu zadań systemów nadzoru wizyjnego.

Rozdział szósty otwiera pierwszą część pracy poświęconą modelom i metodom rozpoznawania aktywności człowieka. Jest w nim omówione rozwiązanie zadania adaptacyjnej, ciągłej kalibracji. Zakłada się, że znane są położenia zbioru punktów znajdujących się na poziomie płaszczyzny powierzchni ziemi. Stosuje się standard dla kamer *Open Network Video Interface Forum* (ONVIF). Metoda ciągłej aktualizacji parametrów kamery jest przetestowana na danych rzeczywistych. Rozdział zamyka opis pierwszej spośród zastosowanych w pracy metod oddzielania obiektów pierwszoplanowych od tła.

W rozdziale siódmym opisano metodę śledzenia i wydzielania obiektów ruchomych. W wykrytych obszarach ruchomych wydzielane są obszary odpowiadające poszczególnym obserwowanym osobom. Wykorzystuje się cechy znalezione metodami ogólnego zastosowania, jak np. SURF i SIFT, natomiast nie korzysta się ze szkieletowej reprezentacji postaci ludzkiej. Analizowane są kontury plam i punkty charakterystyczne na nich znalezione metodą IPAN. W kojarzeniu, łączeniu i dzieleniu obszarów stosuje się metody optymalizacyjne z wykorzystaniem funkcji kosztów, i funkcji różnic (nazywanych przez Autora odległościami), które odzwierciedlają stopień podobieństwa punktów charakterystycznych, oraz szereg zdroworozsądkowych zasad organizujących przeliczanie cech obszarów dzielonych i łączonych. Metody są solidnie przetestowane na danych syntetycznych i rzeczywistych.

Kolejny, ósmy rozdział zawiera opis symbolicznej reprezentacji ruchu, jako zbioru opisów ścieżek poruszających się obiektów. Rozważane jest łączenie ścieżek w klastry, w czym ponownie są użyteczne kolejne zdefiniowane funkcje odzwierciedlające różnice między nimi. Dalej następuje opis wykorzystania opisanych narzędzi do rozpoznawania typu akcji, jak na przykład *chodzenie* i *bieganie*. Prezentowane są wyniki doświadczalnie uzyskanych miar jakości rozpoznawania.

W dziewiątym rozdziale omawia się zagadnienie uczenia dla rozpoznawania akcji człowieka w czasie rzeczywistym. Metoda została wybrana spośród wielu metod znanych z literatury. Niektóre z rozważanych metod były zweryfikowane w konkursach, jakie są organizowane na świecie w tym

zakresie. Dla wybranych jako rokujące nadzieję na zostanie metodą najlepszą przeprowadzono bardzo szeroko zakrojone badania jakości. W tym celu zdefiniowano własną miarę jakości metody, w której uwzględniono zarówno efektywność uzyskiwania wyników, jak i stopień dojrzałości kodu. Autor (wraz z zespołem) przetestował 22 metody na sekwencjach z czterech dużych baz danych sekwencji wideo. Wygrywająca metoda Shi (według pozycji [140] z literatury dysertacji) została poddana szczegółowej analizie, a następnie gruntownie zoptymalizowana i rozszerzona, co pozwoliło nazwać powstałą metodę *rozszerzoną metodą Shi.*

Rozdział 10 rozpoczyna drugą część pracy zatytułowaną *System rozpoznawania aktywności człowieka.* Sam rozdział zawiera opis systemu SAVA od strony architektury. Omawiane są w nim cechy systemu wyróżniające go na tle innych systemów tego typu.

W rozdziale 11 opis systemu SAVA jest kontynuowany. Omawiane są narzędzia wchodzące w skład tego systemu. W szczególności omawia się edytor do anotacji aktywności obiektów (osób) VATRAC oraz symulator tłumu (*Crowd Simulator*) i generator dużych zbiorów danych (*Massive Data Generator*). Narzędzia takie są niezbędne miedzy innymi wskutek wejścia w życie prawa o ochronie danych osobowych (*General Data Protection Regulation GDPR*), co sprawiło, że duże i użyteczne bazy danych o ruchu osób zawierające twarze i nie zanonimizowane stały się bezużyteczne. Omawiane są również dwie bardzo duże anotowane bazy filmów z akcjami ludzi, których współautorem jest Doktorant: VMASS oraz VMASS2.

Rozdział 12 zawiera wnioski z badań oraz wskazania dla możliwych dalszych prac.

W obszernym Dodatku wyjaśniono wybrane zagadnienia organizacji oprogramowania, w szczególności w zakresie śledzenia i wydzielania obiektów poruszających się oraz reprezentacji w metodzie *Bag of Visual Words.* Zamieszczono tam także podręcznik techniczny do systemu SAVA i opis edytora anotacji sekwencji wideo VATRAC.

1.3 Cele pracy

Zacytuję tu cele pracy (rozdział 4.3: *Motivation*) i podejmę próbę ich własnego tłumaczenia.

Objective #1 *There is a method of continuously recognizing multiple individual human actions with the simultaneous spatiotemporal localization of the recognized actions. Component for motion recognition is based on existing single human single action recognition method.*

Cel #1 Istnieje metoda ciągłego rozpoznawania wielu indywidualnych akcji człowieka z jednoczesną lokalizacją czasoprzestrzenną rozpoznanych akcji. Komponent dla rozpoznawania ruchu jest oparty na istniejącej metodzie rozpoznawania *jeden człowiek, jedna akcja.*

Objective #2 *Assuming that it is possible to build a representation of moving objects from a video sequence in the form of symbolic sequences, it is possible to create a general method for creating such representations from a video stream and a method for recognizing motion based on this representation.*

Cel #2 Zakładając, że możliwe jest zbudowanie reprezentacji obiektów ruchomych na podstawie sekwencji wideo w postaci sekwencji symbolicznych, można stworzyć ogólną metodę tworzenia takich reprezentacji na podstawie strumienia wideo, oraz metodę rozpoznawania ruchu opartą na tej reprezentacji.

Objective #3 *A comprehensive framework exists that allows a monitoring system operator with no prior training in computer vision or machine learning to complete the whole process from designing to applying new human action recognition models directly in the monitoring system.*

Cel #3 Istnieje wszechstronna struktura, która pozwala operatorowi systemu monitorującego, nie mającemu uprzedniego szkolenia w zakresie wizji komputerowej ani uczenia maszynowego, ukończyć cały proces, od projektowania do zastosowania nowych modeli rozpoznawania akcji człowieka bezpośrednio w systemie monitorującym.

Jak widać, sformułowanie celów jest równoważne z postawieniem tez pracy.

2 Dyskusja zawartości i wyników rozprawy

2.1 Uwagi pozytywne

2.1.1 Ważność tematyki rozprawy

Często nie jesteśmy zadowoleni z tego, że podlegamy obserwacji, ale w sytuacjach niebezpieczeństwa okazuje się, że ta obserwacja była potrzebna. Łatwo zauważyć, że w miejscach publicznych coraz częściej dochodzi do sytuacji zagrażających naszemu zdrowiu lub nawet życiu. Wtedy szybkość naszej własnej reakcji na zagrożenie i szybkość interwencji odpowiednich służb jest kluczowa. Objętość danych wideo zbieranych stale w wielu miejscach przekracza możliwości, jakie daje bezpośrednia obserwacja przez człowieka. Przeszkodą jest czas, koszty i ryzyko błędów, zarówno przeoczenia jak i niepotrzebnego alarmu. Sytuacje niebezpieczne mogą być różnie zdefiniowane. Inne ryzyka występują na placu zabaw dla dzieci, inne na rynku miejskim, a inne na lotnisku przy kontroli dokumentów. Jeśli na parkingu osiedlowym doszło do uszkodzenia samochodu, to sytuacja do wykrycie jest jeszcze inna, i wolelibyśmy nie stanąć przed koniecznością wymagającego napiętej uwagi, a jednak usypiającego przeglądania tysięcy minut nagrań. Typową reakcją człowieka na zadanie nużące a jednak ważne jest próba jego automatyzacji.

Doktorant uzasadnił potrzebę badań nad analizą strumieni wideo na kilku stronach, w rozdziałach 4.2 i 4.3. Uzasadnienie przeplatane jest wstępną analizą metod, które można zastosować.

2.1.2 Doskonała orientacja w literaturze i wybór metod oparty na testach

Każde działanie Doktoranta zostało poprzedzone bardzo solidną analizą stanu wiedzy na podstawie literatury i innej dostępnej dokumentacji. Po wstępnej selekcji metod z literatury najczęściej przeprowadzano testowanie ich kodów, jeśli były dostępne. Flagowym przykładem jest analiza zawarta w rozdziale 9, gdzie (jak wspomniano wyżej) przetestowano 22 metody na sekwencjach wideo z czterech dużych baz. Testowane metody były uruchamiane na własnym serwerze, więc wyniki można było dowolnie analizować. Jako inny przykład można wskazać przegląd w rozdziale 5.1 *Intelligent Video Analytics – the State of Knowledge*, rozdział 5.2 z podrozdziałami i dyskusję z podsumowaniem w rozdziale 5.3.

2.1.3 Rozszerzenie metody Shi

W przypadku uczenia i rozpoznawania klas akcji człowieka, co stanowi najważniejszą część pracy, metodę, która okazała się najlepsza spośród badanych (metoda Shi według pozycji [140] z literatury dysertacji) rozwinięto i rozszerzono w taki sposób, że używając obliczeń równoległych, również w procesorze graficznym, przyspieszono proces uczenia 3.6-krotnie, a proces rozpoznawania 3.1-krotnie. Przyspieszono również proces optymalizacji wymiaru wektora cech. Do modyfikacji metody zastosowano półautomatyczny proces optymalizacyjny, w którym poszukiwano najlepszej konfiguracji procesów równoległych. Zoptymalizowano również proces analizy zmian sygnału wizyjnego

w czasie. Zbadano kilka koncepcji analizy okien czasowych, gdzie zmieniano liczbę okien, ich częściowe nałożenie lub jego brak, oraz sposób obliczania odpowiedzi z kilku okien: przez głosowanie, przez sumę prawdopodobieństw i przez wybór odpowiedzi z centralnego okna. Jak zrozumiałem z opisu w rozdziałach 9.3 i 9.4, właśnie te optymalizacje dały w efekcie wspomniane przyspieszenia. Pozwoliło to Autorowi nazwać powstałą metodę *rozszerzoną metodą Shi.*

Uważam, że ten fragment pracy jest jej najlepszą częścią ze względu na nowość badawczą. Ma on wprawdzie charakter głownie techniczny, zaś opis jest miejscami trudny do zrozumienia, ale ponieważ mamy w perspektywie doktorat w dyscyplinie *informatyka techniczna (i telekomunikacja)*, to praca zaprezentowana przez Autora pozostaje we właściwym kontekście.

2.1.4 Kalibracja w sposób ciągły

Kamera typu *obrót, pochylenie i zmiana ogniskowej – pan-tilt-zoom* (PTZ) podczas działania zmienia parametry, więc trzeba dokonywać kalibracji w sposób ciągły. Wykorzystanie do tego celu informacji o wybranych punktach, że są na płaszczyźnie ziemi, jest ciekawą koncepcją. Zwykle najniższe punkty obiektów stojących na ziemi to właśnie takie interesujące punkty i łatwo je znaleźć na niemal dowolnym obrazie. Kalibracja obejmuje również nieliniowe zniekształcenia obiektywu.

2.1.5 Brak szkieletowego modelu sylwetki ludzkiej i edytor klas sytuacji

Autor świadomie, na podstawie analizy metod wygrywających we współzawodnictwach, nie zastosował modelu szkieletowego sylwetki ludzkiej, co wielokrotnie podkreśla w tekście. Zamiast tego wykorzystywane są cechy fragmentów obrazu ogólnego zastosowania, jak SURF i SIFT, oraz punkty charakterystyczne na konturach, poddane selekcji (metodą z literatury CMIM [82], gdzie wykorzystuje się maksymalizację informacji wzajemnej). Dzięki temu obiekty wykrywane i śledzone w sekwencjach obrazów nie muszą być osobami ludzkimi. Zatem, również operator edytora akcji nie musi wskazywać anatomicznych części ciała osób, a jedynie nadać odpowiednie deskryptory typu akcji ścieżkom lub zbiorom ścieżek proponowanym przez system.

2.1.6 Inne algorytmy

Doktorant opracował algorytmy na wszystkich kolejnych poziomach przetwarzania w systemie rozpoznawania akcji człowieka – *Human Action Recognition* (HAR). Poza rozszerzoną metodą Shi zasadniczo były to reimplementacje metod znanych z literatury. Metody te były bardzo świadomie wybrane spośród metod znanych, więc były to metody najlepsze według obecnego stanu wiedzy (*state-of-the-art*). Nie było tam znaczącej nowości w sensie naukowym. Jednak trzeba bardzo mocno podkreślić, że w potoku (*pipeline*) przetwarzania strumienia wideo błędy się sumują a dokładności się mnożą, co sprawia, że wymagania dla każdego etapu przetwarzania są bardzo wysokie, jeśli całość ma osiągnąć wymagane parametry jakościowe. Dlatego zapewnienie wysokiej dokładności i bardzo dobrej powtarzalności wyniku każdego etapu przetwarzania ma ogromne znaczenie.

2.1.7 Badania na danych rzeczywistych

Bazy danych, na których badano algorytmy, zawierają obrazy rzeczywiste, ze świata zewnętrznego, bez wstępnej obróbki typu kontrolowane tło, sztuczne oświetlenie, obiekty modelowe. Wyniki testów w postaci miar błędów typu pominięcie i niepotrzebny alarm odpowiadają takim, jakich można się spodziewać w praktyce, a nie w laboratorium.

2.1.8 Ograniczenie potrzeby dobierania parametrów

Autor podkreśla, że automatycznie, za pomocą algorytmów optymalizacyjnych, dobrano parametry w całości procesu przetwarzania w oprogramowaniu SAVA (s. 27, *Contributions*).

2.1.9 Praca z zastosowaniem

Wyniki pracy osiągają poziom przemysłowy i są oferowane na rynku jako aplikacja do użytku w rzeczywistych systemach (konkretnie, w systemie Milestone XProtect^{TM}). Zapewniono takie własności oprogramowania, jak modularność i możliwość wizualizacji wyników pośrednich i końcowych. Dzięki temu użytkownik może zrozumieć, co się dzieje w oprogramowaniu. Zasadniczo, tak lub prawie tak powinno być ze wszystkimi doktoratami w dyscyplinie informatyki technicznej, a w tym przypadku rzeczywiście tak jest.

2.1.10 Bazy anotowanych sekwencji wideo VMASS i VMASS2 i generator sekwencji

Opracowano dwie bardzo obszerne bazy sekwencji wideo ze świata rzeczywistego, z anotacją. Są one większe, niż inne powszechnie dostępne bazy tego typu (na przykład, jest tam ponad 2 miliony zdarzeń, zaś autor wskazuje, że to jest stan „jak dotąd", czyli baza jest powiększana).

Opracowano także generator dużych zbiorów danych tego typu. Potrzeba stosowania sztucznych zbiorów wynika ze wspomnianej już konieczności stosowania się do wymagań ochrony wizerunków osób. Nic jednak nie zastąpi obrazów naturalnych, w których występują wszelkie niedoskonałości prawdziwego świata.

2.1.11 Publikacje

Doktorant zadeklarował 13 publikacji jako zawierających wkład do pracy doktorskiej. Są to publikacje konferencyjne, 12 z konferencji *Asian Conference on Intelligent Information and Database Systems* ACIIDS, w serii LNCS, ranga B w klasyfikacji *Computing Research and Education Association of Australasia* CORE, 20 punktów MEN, jedna z konferencji *International Conference of Numerical Analysis and Applied Mathematics* ICNAAM, w serii AIP Proceedings, brak konferencji w rankingu CORE i spisie MEN, więc 20 punktów jako rozdział w monografii. Doktorant jest obecny na liście autorów na różnych pozycjach, w tym trzy razy na początku i cztery razy na końcu. Jest to dorobek umiarkowany, gdyż brak artykułu w czasopiśmie, ale całkowicie wystarczający.

Doktorant ma razem co najmniej 57 publikacji różnego typu według Google Scholar, indeks h=10; według ResearchGate 53, h=10, według Scopus 47, h=7, według Web of Science 41, h=7. Jest to bardzo dobry wynik.

2.2 Uwagi dyskusyjne i krytyczne

2.2.1 Uwagi merytoryczne

Co zrobić jeśli poziom ziemi nie jest płaszczyzną? Wiele placów miejskich ma powierzchnię ukształtowaną tak, żeby deszczówka dobrze ściekała. Zdarzają się także znaczne odstępstwa od płaskości spowodowane zużyciem, uszkodzeniami, lub złą jakością. Czy ta prozaiczna okoliczność nie uniemożliwia ciągłej kalibracji kamery, opisanej w pracy?

Optymalizacja parametrów Tu chciałbym zapytać Autora, jak ocenia stopień, w jakim udało się osiągnąć optimum globalne dla bardzo wielu parametrów, które są w algorytmach

oprogramowania SAVA. Wszystkie metody odległościowe wymagają parametrów, nawet same funkcje niepodobieństwa (nie muszą to być odległości w ścisłym sensie) mają wagi dla niepodobieństwa różnych wielkości, jak choćby odległość w przestrzeni i różnica w czasie. Są liczne progi dla tych funkcji. Wreszcie, procesy uczenia polegają na ustalaniu wartości bardzo licznych parametrów, choć tu właśnie proces uczenia zapewnia osiągnięcie optimum, które ma wartość o ile jest stabilne względem nieistotnej zmienności danych.

Zniekształcenia kątowe Nie są one uwzględnione w wyjściowym wzorze (6.1), więc nie mogą się pojawić dalej. Wprawdzie parametr s_x określa się na str. 63[8] jako odpowiedzialny za zniekształcenia styczne (*tangential*), to jednak jest to w rzeczywistości tylko odchylenie od kwadratowości piksla. Można pominąć zniekształcenia kątowe, ale trzeba to napisać.

2.2.2 Uwagi pośrednie

Uwagi podane tutaj są na pograniczu uwag merytorycznych i redakcyjnych. Nie mają więc dużego krytycznego ciężaru merytorycznego. Nie wpływają one na moją ogólną wysoką ocenę rozprawy doktorskiej.

S. 65, wzory (8.1), (8.2): Jaki jest cel dodawania 1 w nawiasie kwadratowym w (8.1)? Czy jedynka nie powinna raczej być dodana do argumentu pod logarytmem, żeby wynik nie był ujemny? Brak wyjaśnienia oznaczeń A_s, B_s oraz B. Dalej, w tekście "translated in space by (x,t) and in time by t" powinno być (x,y).

S. 131, rozdział 8.1: Zdefiniowano funkcje podobieństwa (a właściwie niepodobieństwa; nie muszą one mieć własności odległości), opisano klasteryzację deskryptorów, natomiast właściwie nie zdefiniowano czym jest deskryptor. Można się łatwo domyślić, ale przecież nie o to chodzi.

S. 133, wzory (8.7), (8.8): Jaki jest cel odejmowania 1 w ostaniem nawiasie?

S. 137, wzory (8.13): Czułość to $SE = \frac{TP}{TP+FN}$ (w mianowniku są wszystkie obiekty rzeczywiście pozytywne), a nie $\frac{TP}{TN+FN}$. Podobnie specyficzność to $SP = \frac{TN}{TN+FP}$ (w mianowniku są wszystkie obiekty rzeczywiście negatywne), a nie $\frac{TN}{TN+FN}$. Mam nadzieję, że w oprogramowaniu jest dobrze.

S. 139, reason 2: Wybrano tylko klasy *chodzenie* i *bieganie*, na rynku w Gliwicach. Tak się składa, że biegają zwykle dzieci, nie dorośli. Czy ich systematycznie niższy wzrost nie wpływa nadmiernie na poprawienie wyników klasyfikacji?

S. 151, wzór (9.1): Wynik silnie zależy od współczynnika przy zmiennej CML_k, czyli 5. Niewielka zmiana tego współczynnika zmieni kolejność metod w tabeli 9.4 i metoda Shi może nie być najlepsza. Skąd wartość 5? Czy nie lepiej od razu zdecydować, że najpierw wybieramy metody na najwyższym poziomie dojrzałości kodu, a spośród nich (choć jest tylko jedna) wybieramy tę, która ma największą wartość wydajności? Albo badamy wydajność, a potem bierzemy pod uwagę dojrzałość kodu (i chyba znów Shi wygrywa)?

2.3 Uwagi redakcyjne i techniczne

Pod względem redakcyjnym praca jest opracowana prawidłowo, a miejscami nawet bardzo starannie. Jednak znalazło się w niej bardzo wiele błędów i potknięć, tak redakcyjnych jak i językowych.

Poniżej dla przykładu podam niektóre błędy. Powtarzają się zdania ze złą strukturą, braki przyimków, nieuzgodniona liczba lub osoba, powtórzone wyrazy, braki odwołań do numerów tabel czy rysunków. Błędy są uporządkowane według stron, a nie według ważności. Błędy te nie mają dużego wpływu na merytoryczną wartość rozprawy. Teksty dodane są oznaczone kolorem zielonym i podkreśleniem, a teksty usunięte ~~kolorem czerwonym i przekreśleniem~~.

S. 1^{6}: PTZ: nawet, jeśli w pracy jest spis skrótów, to każdy skrót trzeba wyjaśnić przy pierwszym użyciu. Szczególnie dotyczy to skrótu pracy, ponieważ powinien on być zrozumiały w oderwaniu od całej treści pracy.

S. 1^{12-13}: "...and continuous tracking of recognized ~~in video stream~~ moving objects in video stream". Język angielski ma zwykle stałą strukturę zdania: podmiot, orzeczenie, dopełnienie bliższe, dopełnienie dalsze.

S. 3_{12}: "bezszkieletowej".

S. 5_{9-8}: "A considerable amount of work ~~took the completion~~was required for the completion of the methods which were published, but were incomplete, not unified, not verified on known benchmarks, ~~without~~and had no implementation~~but published methods~~.".

S. 6^{7-8}: "~~sympathetic~~friendly". *Sympathetic* znaczy przede wszystkim *wykazujący zrozumienie, współczujący*, a nawet *ubolewający*. Dopiero na dalszych miejscach jest *miły, sympatyczny, życzliwy*.

S. 12_{13}: "*Movement* is defined as a pose change...".

S. 12_{11}: "~~Thomas B.~~ Moeslund et al. [186] use...". Zawsze podajemy tylko nazwisko.

S. 13, top: Terminologia: *person* czy *actor*?

S. 13, Definition 4.1: Zdefiniowana jest akcja, a właściwie podane są jej ograniczenia. Następnie mowa jest o rozpoznawaniu akcji oraz o strumieniu wideo (przy okazji, T jest skalarem i nie ma sensu pisać $\mathbf{T}$, bo to już zmienna inna niż T). Dalej mowa jest o więzach przestrzennych. To już cztery definicje.

Zdanie pierwsze jest trudno zrozumiałe, proponuję "*Action* **A** is a full body movement of one person, constrained ~~by the~~in time ~~a~~and distinguished in space and time ~~full body movement of one person~~, in the video stream". Jest jasne, że to nie osoba, a ruch jest ograniczony w czasie, natomiast struktura zdania jest lepsza.

S. 18, *Objective #2*: "...creating such representations from a video stream and a method for recognizing motion based on ~~this representation~~these representations.". Both equally plural.

S. 19, podpis Rys. 4.2 i dalszych rysunków: "(`Source: Own elaboration`)". Lepiej jest przy pierwszym rysunku napisać: "All the images, except where stated otherwise, are the own works of the author.", i nie powtarzać już dalej tego tekstu. Jednak, jeśli Uczelnia ma inne wymagania, to trzeba się do nich stosować.

S. 21_{12}: "...KTH ~~[229]~~ dataset [229]...".

S. 29, tekst pod podpisem rysunku 4.3: "The work is organized as ~~depicted~~depicted in Figure 4.3. In Chapter 5, we briefly present...". Czytelnik spodziewa się, że na Rys. 4.3 są zaznaczone numery rozdziałów, a tak nie jest.

Dalej: "*Estimation, Tracking, Deep Le~~r~~arning,...*".

S. 29$_{3-2}$: "The idea of temporal action localization (...) ~~shows~~is shown in section 9.3, respectively.". Jak wiadomo, język angielski lubi stronę bierną.

S. 45$_{8-7}$: "The main elements of ViT ~~presents~~is presented in Fig 5.6.".

S. 54, bottom: "An extended variant of this method ~~presents~~is presented in chapter 8. The idea of temporal action localization in a video stream ~~we propose~~is proposed by us in section 9.3. The method for continuous calibration with tracking of camera pose, also ~~utilizing~~used for spatial localization, ~~presents~~is presented in section 6, respectively.".

S. 61^{8}: "The ~~rest~~remaining commands of this kind send back messages...".

S. 63$_{4}$: "~~P_{3x3}~~$P_{3\times3}$". To przykład powrotu do archaicznej metody symulowania znaków matematycznych literami. A przecież LaTeX ma tak wielkie możliwości.

S. 67, podpis Rys. 6.2: "Radial grid..." Siatka nie jest radialna, lecz raczej liniowa, zdeformowana.

S. 70, podpis Rys. 6.3: "As one can see..." Jednak tego nie widać ani w pracy wydrukowanej, ani w wersji elektronicznej, gdzie rysunki są mocno skompresowane, stratnie.

S. 70, tytuł rozdziału i dalsza treść: "Navigation is a real-time camera calibration procedure..." *Navigation* to jednak zawsze *nawigacja, żegluga.* Skąd taka nazwa?

S. 107, wzór (7.17) i w innych miejscach: Oznaczenia takie, jak bx, bxx powinny raczej wyglądać tak: b_x, b_{xx}.

S. 132^{6}: "The approximate solution can be obtained iteratively...".

S. 133$_{6}$: "...will be used by ~~used by~~ the normalized distance...".

S. 135^{1}: "~~Relations between motion steps and relational descriptors The aim of motion descriptions...~~

Relations between motion steps and relational descriptors The aim of motion descriptions...".

Here, I used `{\bf <text>~~}`, but one can use `\paragraph{}`.

S. 135^{4}: "~~...the distance is infinity. Constructing relational symbolic features Proceed separately for each...~~

...the distance is infinity.

Constructing relational symbolic features Proceed separately for each...".

Same as above.

S. 139$_{4}$: "...used to ~~validate~~test the model.". Co innego walidacja, co innego testowanie. Tu było testowanie, jak zresztą napisano na następnej stronie, linia 15 od dołu.

S. 140$_{15-14}$**:** "The ~~following~~ classification results shown in Table 8.1 were obtained~~:~~. As can be seen,...".

S. 153[8]**:** "...and the rule ~~9.1~~(9.1)". To jest równanie, więc numer podajemy w nawiasach.

S. 169[10]**:** "...sets. ~~Allows~~It allows for simultaneous annotation...".

W całej pracy: Nazwy funkcji, dla odróżnienia od nazw zmiennych, są pisane czcionką prostą jeśli nie są jednoliterowe. Przykłady: sin, ln. Dlatego takie funkcje, jak Dlog, NDlog powinny być pisane właśnie tak. Natomiast zapis $NDlog(\cdot)$ można równie dobrze rozumieć jako iloczyn zmiennej N i funkcji Dlog$(\cdot)$, a nawet zmiennych N, D i funkcji log$(\cdot)$.

Podobnie ~~*symbol(A)*~~symbol(A), ~~*Assign(·)*~~Assign$(\cdot)$, ~~s_{prev}, s_{next}~~$s_{\text{prev}}, s_{\text{next}}, s_{\text{max}}^{\text{merge}}$. Jest z tym niestety dużo roboty: s_{next} pisze się `$s_\textrm{\scriptsize next}$`. Można sobie zdefiniować odpowiednie makro w LaTeXu. Nawet takie oznaczenia, jak TP, FP, ... i podobne pisze się `$\textit{TP, FP}$`, a nie TP, FP czyli `$TP, FP$`. Widać różnicę w odstępach, widać, że to nie iloczyn zmiennych T i P.

Zdarzają się teksty wysunięte poza prawy margines, np. s. 14^{11}, $40_{10,9}$, 40_2. W LaTeXu nazywa się to *overfull* i jest raportowane w podsumowaniu kompilacji jako *bad boxes: <number>*. Jeśli *<number>* $\neq 0$, to trzeba poszukać w logu. Zwykle edytory w tym pomagają.

3 Podsumowanie

W rozprawie można wyróżnić istotne elementy oryginalne oraz inne pozytywne aspekty omówione w rozdziale 2.1, zaś uwagi dyskusyjne i krytyczne omówione w rozdziale 2.2 nie umniejszają w istotny sposób wartości pracy.

Cele pracy zostały osiągnięte.

W rozprawie wszechstronnie opracowano metodę rozwiązania zadania rozpoznawania akcji osób i grup osób. Wymaga to oczywiście uprzedniej detekcji osób i znalezienia ich ścieżek oraz ewentualnego ich pogrupowania. Wśród opracowanych metod wyróżnia się metoda uczenia i rozpoznawania klas akcji człowieka, która została nazwana *rozszerzoną metodą Shi*. Zaproponowane rozszerzenia są tak daleko idące, że można je uważać za znaczące osiągnięcie mające cechy nowości naukowej.

Wniosek o wyróżnienie rozprawy Ze względu na opracowanie *rozszerzonej metody Shi*, oraz inne pozytywne strony pracy wymienione wyżej, w szczególności bardzo dobrą znajomość literatury, umiejętność wiarygodnego testowania metod, oraz kompletność i wysoki poziom zaawansowania technologicznego wytworzonego oprogramowania, **stawiam wniosek o wyróżnienie rozprawy**.

Wnioski końcowe Powyższy opis, uwzględniający uwagi pozytywne jak również uwagi dyskusyjne i krytyczne, uzasadnia moje ostateczne wnioski o następującej treści.

Recenzowana rozprawa w postaci opracowania pisemnego **stanowi oryginalne rozwiązanie problemu naukowego** w dyscyplinie *informatyka techniczna i telekomunikacja*. Rozprawa **prezentuje ogólną wiedzę teoretyczną kandydata w tej dyscyplinie oraz umiejętność samodzielnego prowadzenia pracy naukowej**. Zagadnienie badawcze zostało prawidłowo postawione i skutecznie rozwiązane, a rozwiązanie zostało rzetelnie zweryfikowane. Tym samym rozprawa spełnia wymagania obowiązującego prawa w zakresie rozpraw doktorskich. Rozprawę oceniam pozytywnie i stawiam wniosek o skierowanie jej do dalszych etapów przewodu doktorskiego.