Politechnika Wrocławska
Wydział Informatyki i Zarządzania
Instytut Informatyki

Rozprawa doktorska

Modelowanie generujące w zastosowaniu do śledzenia ruchu człowieka

Adam Gonczarek

Promotor: Prof. dr hab. inż. Jerzy Świątek

Wrocław 2013

Podziękowania

Na wstępie chciałbym podziękować mojemu promotorowi, Panu Profesorowi Jerzemu Świątkowi, za pomoc przy realizacji pracy, wszelkie uwagi i sugestie, a także za sprawowanie opieki naukowej od momentu pisania pracy magisterskiej.

Ponadto chciałbym podziękować moim trzem serdecznym przyjaciołom, którzy w istotnym stopniu przyczynili się do ostatecznej formy tej rozprawy. Wymieniając w kolejności alfabetycznej, Pawłowi Siemionko, najwybitniejszemu programiście, którego miałem kiedykolwiek przyjemność spotkać, za wszelką pomoc przy pisaniu kodu oraz wiele trafnych sugestii, Jakubowi M. Tomczakowi za wspólne zgłębianie tajników uczenia maszynowego, cenne uwagi, motywowanie do pracy i ciągłe podnoszenie poprzeczki, oraz Michałowi Walczakowi za godziny dyskusji o charakterze interdyscyplinarnym, niezliczoną ilość błyskotliwych spostrzeżeń i wskazanie jedynego słusznego punktu widzenia na wiele spraw.

Na koniec chciałbym podziękować moim Rodzicom, którzy przez cały czas mnie wspierali i dopingowali w pisaniu tej rozprawy.

Pracę dedykuję mojemu dziadkowi
– Ś. P. Czesławowi Gonczarkowi.

Część niniejszej rozprawy została wykonana w ramach stypendium współfinansowanego przez Unię Europejską w ramach projektu „Rozwój potencjału dydaktyczno-naukowego młodej kadry akademickiej Politechniki Wrocławskiej”.

Spis treści

Podziękowania **ii**

Spis treści **iii**

1 Wstęp **1**

1.1 Wprowadzenie 1

1.2 Aktualny stan badań 4

1.3 Teza, cel i zakres pracy 8

1.4 Układ pracy 10

2 Koncepcja systemu śledzącego **12**

2.1 Reprezentacja ciała człowieka 12

2.1.1 Obroty w przestrzeni trójwymiarowej 13

2.1.2 Drzewo kinematyczne 20

2.1.3 Wektor stanu 24

2.2 System pomiarowy 25

2.2.1 Kalibracja kamer 26

2.2.2 Projekcja perspektywiczna 29

3 Sformułowanie problemu **31**

3.1 Problem estymacji pozy 31

3.2 Problem śledzenia ruchu człowieka 34

3.2.1 Dynamika w pobliżu niskowymiarowej rozmaitości 38

3.3 Ocena jakości rozwiązania 41

4 Metody filtrowania **43**
4.1 Filtr cząsteczkowy 43
4.2 Wyżarzany filtr cząsteczkowy 50
4.3 Filtr cząsteczkowy uwzględniający niskowymiarową rozmaitość 53

5 Modele wiarygodności **60**
5.1 Modelowanie ciała 60
5.1.1 Mapa głębi 61
5.2 Model oparty na sylwetkach 63
5.2.1 Problem oddzielania tła 63
5.2.2 Funkcja wiarygodności 66
5.3 Model oparty na krawędziach 68
5.3.1 Mapa krawędzi 69
5.3.2 Funkcja wiarygodności 71
5.4 Model oparty na lokalnych deskryptorach 72
5.4.1 Lokalne deskryptory 74
5.4.2 Model wyglądu 77
5.4.3 Funkcja wiarygodności 80
5.5 Łączenie modeli wiarygodności 82

6 Modele dynamiki **83**
6.1 Model podstawowy 83
6.2 Model uwzględniający strukturę rozmaitości 85
6.2.1 Odtworzenie struktury rozmaitości 87
6.2.2 Dynamika na rozmaitości 94
6.2.3 Dynamika w przestrzeni stanów z uwzględnieniem rozmaitości . . . 95

7 Badania empiryczne **98**
7.1 Zbiór danych 98
7.2 Badanie jakości śledzenia ruchu z użyciem filtra cząsteczkowego uwzględniającego niskowymiarową rozmaitość 100
7.3 Badanie jakości śledzenia ruchu z użyciem modelu wiarygodności opartego na lokalnych deskryptorach 112

8 Podsumowanie i uwagi końcowe **124**
8.1 Oryginalny wkład pracy w dziedzinę śledzenia ruchu człowieka 124
8.2 Kierunki dalszych badań 125

Bibliografia **127**

Spis symboli i skrótów **144**

Spis rysunków **153**

Spis tabel **156**

Skorowidz **157**

Rozdział 1

Wstęp

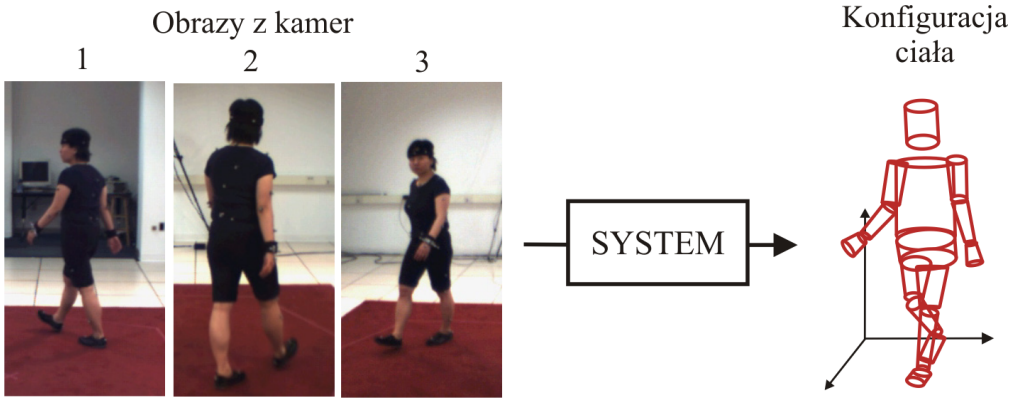
1.1 Wprowadzenie

Śledzenie ruchu człowieka (ang. human motion tracking) polega na sekwencyjnym odtwarzaniu konfiguracji ciała i stanowi ważny praktyczny problem. Obecnie podstawową aparaturą pozwalającą na dokładną realizację tego zadania są *systemy MOCAP*[1], które przy pomocy specjalnych znaczników (ang. markers) rozmieszczonych w różnych punktach na ciele umożliwiają rejestrowanie położenia i orientacji poszczególnych kończyn. Systemy te są powszechnie stosowane w przemysłach filmowym i gier komputerowych, gdzie wykorzystuje się je do tworzenia realistycznych animowanych postaci, bazując na sposobie poruszania się rzeczywistych ludzi. Istnieją jednak powody, które wykluczają *systemy MOCAP* z powszechnego użycia. Po pierwsze rozmieszczenie znaczników na ciele musi być wykonane w sposób precyzyjny, wymaga czasochłonnego przygotowania i kalibracji, człowiek powinien być ubrany w strój, który zapobiega ich przesuwaniu, a cały proces śledzenia musi się odbywać w dostosowanym pomieszczeniu, gdzie rozstawione są urządzenia rejestrujące pomiary z poszczególnych markerów. Ogranicza to ich zastosowanie jedynie do warunków studyjnych. Po drugie koszt tych systemów jest bardzo wysoki i waha się od kilku do kilkudziesięciu tysięcy euro.

W związku z tym obecnie prowadzone są intensywne badania nad stworzeniem tzw.

[1] System MOCAP jest skrótem od Motion Capture i oznacza system do rejestrowania trajektorii poruszającego się człowieka lub innego obiektu.

bezznacznikowych (ang. markerless) systemów do *śledzenia ruchu człowieka*. Ogólna ich idea polega na odtworzeniu konfiguracji na postawie obserwacji z jednej lub kilku kamer, bez użycia jakichkolwiek markerów czy dodatkowych urządzeń (rysunek 1.1).



Rysunek 1.1: Idea systemu do bezznacznikowego odtwarzania konfiguracji ciała człowieka.

Motywacją do pracy nad tym problem są liczne potencjalne zastosowania: począwszy od wspomnianych systemów do rejestrowania trajektorii, przez systemy interakcji człowiek-komputer, pozwalające na sterowanie aplikacjami przy pomocy gestów, systemy do nauki tańca i sztuk walki, systemy do szkolenia sportowców i do diagnostyki medycznej, które dają możliwość oceny poprawności wykonywanych sekwencji ruchów, aż po systemy wspomagające monitoring i pozwalające na wykrywanie nietypowych zachowań w miejscach użyteczności publicznej (np. na lotniskach) oraz na identyfikację tożsamości na podstawie sposobu poruszania się [79, 110, 124].

Pomimo dużego postępu w tym obszarze i pierwszych komercyjnych sukcesów[2], problem bezznacznikowego *śledzenia ruchu człowieka* pozostaje w zasadniczej części nierozwiązany. Wśród najważniejszych czynników, które stanowią o jego stopniu skomplikowania należy wymienić:

1. Zmienne warunki oświetlenia, które wpływają na jakość obrazów uzyskiwanych z

[2]W 2010 roku firma Microsoft wprowadziła na rynek czujnik Xbox Kinect, które pozwala na sterowanie grami komputerowymi przy pomocy gestów i ruchów ciała. Urządzenie wykonuje pomiary z kamery podczerwonej w oparciu o strukturalne światło wyświetlane na człowieku, przez co jego zastosowanie ograniczone jest jedynie do niewielkich zamkniętych pomieszczeń.

kamer, a także powodują, że jednej konfiguracji ciała może odpowiadać cała gama obrazów o różnym nasyceniu pikseli.

2. Zmienny wygląd człowieka, który wynika z rodzaju i koloru ubrania, rodzaju fryzury itp. Ponownie tej samej konfiguracji może odpowiadać wiele różnych obrazów.

3. Różnorodny i często ruchomy charakter otocznia śledzonej postaci, który wprowadza wiele nadmiarowej informacji do pomiarów, na którą system powinien być niewrażliwy.

4. Częściowe przesłonięcie niektórych fragmentów ciała przez inne fragmenty lub elementy otoczenia. Powoduje to brak dostatecznej ilości informacji do jednoznacznego odtworzenia konfiguracji.

5. Próba wnioskowania o trójwymiarowej rzeczywistości na podstawie dwuwymiarowych obrazów i ponowny brak dostatecznej ilości informacji do jednoznaczności w odtwarzaniu ułożenia ciała.

6. Wysokowymiarowy charakter wektora opisującego konfigurację człowieka, a w konsekwencji trudności z przeszukiwaniem przestrzeni potencjalnych konfiguracji.

Powyższe problemy lokalizują zagadnienie *śledzenia ruchu człowieka* na styku dwóch dziedzin informatyki – *widzenia komputerowego* (ang. computer vision) [55, 138] i *uczenia maszynowego* (ang. machine learning) [17, 68, 106, 112]. W szczególności problemy 1. i 2. wymagają m. in. zaawansowanych technik ekstrakcji cech z obrazów [3, 12, 14, 39, 100, 105, 137], a także złożonych metod detekcji poszczególnych części ciała [51, 52, 167]. Do problemu 3. wykorzystuje się m. in. segmentację [37, 171] i *oddzielanie tła* [31, 49, 121, 145]. Problemy 4. i 5. wymagają użycia technik wnioskowania probabilistycznego [75, 80, 104, 117, 147] lub metod predykcji dla wielomodalnych rozkładów [76, 127]. Problem 6. wymusza użycie metod pozwalających zamodelować strukturę wysokowymiarowej przestrzeni konfiguracji, w szczególności technik *klasteryzacji* [42, 61, 64, 69, 81] i *redukcji wymiarów* [11, 18, 90, 107, 133, 154, 170].

Na koniec warto zaznaczyć, jak *śledzenie ruchu człowieka* zlokalizowane jest pośród innych problemów rozważanych w literaturze. Po pierwsze jest ono sekwencyjną wersją

zagadnienia *estymacji pozy* (ang. pose estimation), gdzie odtwarzana jest konfiguracja ciała na podstawie pojedynczego zdjęcia lub kilku zdjęć z różnych perspektyw. Po drugie jest ono szczególnym przypadkiem śledzenia obiektów przegubowo połączonych (ang. articulated object tracking), którego innym przykładem jest zadanie śledzenia dłoni (ang. hand tracking) [50, 103, 111, 146]. Po trzecie stanowi ono punkt wyjścia dla zagadnienia *rozpoznawania akcji* (ang. action recognition) [20, 28, 126, 173], które polega na klasyfikowaniu fragmentów trajektorii ruchu do klas określających rodzaje zachowań.

1.2 Aktualny stan badań

Problemy *śledzenia ruchu człowieka* i *estymacji pozy* są zagadnieniami szeroko rozpatrywanymi w literaturze. Istnieje wiele sposobów podziału tworzonych metod. Do najważniejszych można zaliczyć podział ze względu na liczbę użytych kamer, gdzie wyróżniamy techniki bazujące na obrazach z jednej perspektywy (ang. monocular) i z wielu kamer (ang. multiview), podział ze względu na reprezentację konfiguracji ciała – pozy dwuwymiarowe i trójwymiarowe, a także podział ze względu na sposób modelowania zależności pomiędzy obserwacjami z kamer i szacowaną konfiguracją ciała – podejście *dyskryminacyjne* (ang. discriminative) i podejście *generujące* (ang. generative). Do scharakteryzowania istniejących algorytmów w pracy został wykorzystany ostatni podział. Szczegółowy opis istniejących metod według różnych zasad podziału można znaleźć w pracach przeglądowych [79, 110, 124].

Podejście *dyskryminacyjne* charakteryzuje się tym, że bezpośrednio modelowana jest zależność pomiędzy obserwacjami z kamery i konfiguracją ciała, tj. zadaje się postać warunkowego rozkładu prawdopodobieństwa na konfigurację pod warunkiem obserwacji. Z obrazów dokonuje się ekstrakcji cech, by możliwie zredukować ich różnorodność dla tej samej konfiguracji ciała. Wykorzystuje się do tego zarówno podstawowe techniki, jak wyodrębnienie *sylwetek* [4, 48, 108], jak również bardziej zaawansowane deskryptory [19, 83]. Następnie szacowana jest konfiguracja ciała w oparciu o decyzję podjętą na podstawie nauczonego modelu, który jako zmienne wejściowe przyjmuje cechy uzyskane z obrazu. Pierwsze prace wykorzystywały proste modele, jak *regresję liniową* z *regularyzacją* L2 (ang. ridge regression) czy *Support Vector Machine* [4], następnie stosowano modele pozwalające

na dopasowanie wielomodalnych rozkładów prawdopodobieństwa, jak regresja z użyciem bliźniaczych *procesów Gaussa* [19] czy *mieszanina ekspertów* (ang. mixture of experts) [83]. Najnowszy trend wykorzystuje modele ze współdzieloną przestrzenią (ang. shared-space models), które zakładają, że konfiguracje ciała i obrazy determinowane są przez wspólny nieobserwowany czynnik – zmienną ukrytą (ang. latent variable)[3]. Wyróżnić można wśród nich można podejścia stosujące model *Shared Gaussian Process Latent Variable Model* (sGPLVM) [48] oraz model *Shared Kernel Information Embedding* (sKIE) [108]. *Modele dyskryminacyjne* charakteryzują się wysoką skutecznością nawet dla obserwacji z pojedynczej kamery, jednakże tylko w przypadku, gdy dane ustawienie ciała wchodziło w skład ciągu treningowego, na podstawie którego model był nauczony. Zazwyczaj oznacza to, że modele te są mocno zawężone do wybranych konfiguracji i źle uogólniają dla nowych, niewidzianych dotąd obserwacji. Problem ten można częściowo eliminować stosując *uczenie z częściowym nadzorem* (ang. semi-supervised learning), gdzie do zbioru uczącego wprowadza się dużą ilość danych wejściowych, dla których nie znamy odpowiadającego im wyjścia [83].

W przeciwieństwie do *modeli dyskryminacyjnych*, podejście *generujące* zakłada, że zależność pomiędzy obserwacjami i konfiguracją ciała modelowana jest w sposób pośredni, tj. modeluje się osobno rozkład a priori na przestrzeń konfiguracji oraz funkcję wiarygodności, która jest warunkowym rozkładem na obserwację przy ustalonej konfiguracji ciała. W tej grupie wyróżnić należy dwa główne nurty, które różnią się przyjętą reprezentacją ciała człowieka. Pierwszy z nich bazuje na tzw. *drzewie kinematycznym* (ang. kinematic tree), które zakłada, że konfiguracja człowieka jest wyrażana jako jedna spójna całość. Drugi natomiast wykorzystuje *modele oparte na częściach* (ang. part-based model), gdzie ułożenie każdej części ciała rozpatrywane jest osobno i korygowane na podstawie wiedzy apriorycznej.

Wśród podejść wykorzystujących *drzewo kinematyczne* możemy wskazać dwie techniki wnioskowania. Pierwsza z nich polega na znalezieniu konfiguracji ciała będącej estymato-

[3]Zważywszy na fakt, że modelowany jest łączny rozkład cech z obrazów i konfiguracji ciała pod warunkiem zmiennej ukrytej, to model ze współdzieloną przestrzenią może być traktowany również jako *model generujący*. W pracy zalicza się go do *modeli dyskryminacyjnych* ze względu na sposób wnioskowania z jego użyciem, tj. najpierw wyznacza się zmienną ukrytą na podstawie obserwacji, co można traktować jako dodatkową ekstrakcję cech, a następnie na jej podstawie estymowana jest konfiguracja ciała.

rem maksymalnego prawdopodobieństwa a posteriori (MAP), z użyciem wybranej metody optymalizacji, jak standardowe metody gradientowe [161, 162], modyfikacje *symulowanego wyżarzania* [59], algorytm *Stochastic Meta Descent* (SMD) [84] czy algorytmy genetyczne [177]. Druga grupa metod szacuje cały rozkład a posteriori i na jego podstawie wyznacza estymaty konfiguracji ciała. Zazwyczaj odbywa się to z użyciem *filtra cząsteczkowego* (ang. particle filter) [75] i jego modyfikacji, np. *wyżarzanego filtra cząsteczkowego* [44]. Niemniej istnieją również inne podejścia, jak wykorzystanie *rozszerzonego filtra Kalmana* (ang. extended Kalman filter) [109] czy metody *Covarianced Scaled Sampling* [144].

Rozkład a priori na konfigurację ciała standardowo przybliżany jest poprzez dyskretną aproksymację w jednym lub wielu punktach, aczkolwiek ze względu na wysoki wymiar przestrzeni potencjalnych ułożeń ciała stosuje się dodatkowe modele poprawiające wiedzę aprioryczną. Począwszy od prostych ograniczeń na zakres ruchomości poszczególnych stawów [140], poprzez modele korzystające z praw fizyki do generowania potencjalnych ułożeń kończyn [24], do modeli przybliżających *rozmaitość* (ang. manifold), na której rozłożone są rzeczywiste konfiguracje ciała i sposób poruszania się po niej. Ostatnia grupa wykorzystuje metody *uczenia bez nadzoru* (ang. unsupervised learning), gdzie wśród technik *klasteryzacji* wykorzystuje się m. in. *mieszaniny rozkładów Gaussa* (ang. mixture of Gaussians) [72], model *Hierarchical Hidden Markov Model* [120], model *Variable Length Markov Model* [26, 71]. Z technik *redukcji wymiarów* używa się np. *analizy głównych składowych* (ang. principal component analysis) [162], modelu *Gaussian Process Latent Variable Model* i jego modyfikacji [71, 155, 161], jak również kodowania *rozmaitości* przy pomocy zbioru zmiennych binarnych z użyciem *ograniczonej maszyny Boltzmanna* (ang. restricted Boltzmann machine) [153]. Stosuje się także modele, które z założenia są połączeniem *klasteryzacji* i *redukcji wymiarów*, jak *mieszanina analiz czynnikowych* (ang. mixture of factor analyzers) [97].

Wnioskowanie odbywa się poprzez połączenie rozkładu a priori z *modelem wiarygodności*, który poprawia wiedzę o potencjalnej konfiguracji ciała poprzez uwzględnienie informacji wynikającej z obserwacji bieżących obrazów z kamer. W *modelach generujących* wykorzystujących *drzewo kinematyczne* do wnioskowania stosuje się tzw. *podejście top-down*, tj. najpierw generowana jest potencjalna konfiguracja (w wyniku dyskretnego przybliżenia rozkładu a priori lub w kolejnych krokach numerycznej optymalizacji), a następnie ocenia

się jej jakość poprzez funkcję wiarygodności. Zazwyczaj wykorzystuje się dodatkowy *model ciała* (ang. body model), który w odpowiednim ustawieniu rzutowany jest na obraz z danej perspektywy i liczona jest różnica pomiędzy nim i rzeczywistą obserwacją. Niemniej są podejścia, które nie korzystają z *modelu ciała* i od razu generują obraz do porównania poprzez próbkowanie ze wspólnej *rozmaitości* dla konfiguracji i obrazów [66, 77, 94]. Niezależnie od podejścia *model wiarygodności* musi pozwalać na porównanie rzeczywistego i hipotetycznego obrazu. Standardowo stosuje się w tym celu binarne mapy przedstawiające obrys śledzonej postaci – tzw. *sylwetki* (ang. silhouette), często wspomagane przez dodatkową mapę krawędzi (ang. edge map) [32, 44, 140]. Należy jednak zaznaczyć, że w literaturze spotyka się również inne *modele wiarygodności*, m. in. bazujące na kolorach [130], na punktach odniesienia [128], na *przepływie optycznym* (ang. optical flow) [24, 144], a także na porównaniu rozkładów trójwymiarowych wokseli [26].

Drugą grupę metod bazującą na podejściu *generującym* stanowią techniki wykorzystujące *modele oparte na częściach*, gdzie każdy element ciała rozpatrywany jest osobno. W tym przypadku wnioskowanie praktycznie zawsze polega na szukanie estymatora MAP i w przeciwieństwie do metod stosujących *drzewo kinematyczne* wykorzystywane jest *podejście bottom-up*, tj. najpierw wykrywane są poszczególne części ciała przy pomocy dedykowanych detektorów (z dokładnością do prawdopodobieństwa), a następnie uzyskana w ten sposób informacja jest ze sobą składana w wybranej procedurze wnioskowania[4].

Modele oparte na częściach mogą być traktowane jako szczególne przypadki *markowskich pól losowych* (ang. Markov random field), w których poszczególne zmienne losowe oznaczają położenie części oraz wyróżnia się dwa rodzaje funkcji potencjału: pierwsza związana jest z wyglądem każdego z fragmentów, a druga z ich wzajemnymi relacjami, jak np. położenie względem siebie. Najczęściej stosowanym modelem jest *struktura obrazkowa* (ang. pictorial structure) [54], która zakłada, że funkcje wzajemnego potencjału są gaussowskie i dodatkowo sieć połączeń w modelu ma strukturę drzewa. Zainteresowanie tymi modelami wzrosło dzięki pracom [51, 52], w których zaprezentowano efektywną procedurę do wyliczania estymatorów MAP opartą na *programowaniu dynamicznym*, która ma liniową złożoność względem liczby pikseli na obrazie dzięki zastosowaniu tzw. transfor-

[4] Są wyjątki od tej reguły i zdarzają się prace, gdzie aproksymuje się cały rozkład a posteriori i korzysta z *podejścia top-down* [41].

macji odległości (ang. distance transform) [53], która łączy informację o wyglądzie części i wzajemnych potencjałach w postaci gaussowskiej. *Struktury obrazkowe* są powszechnie stosowane do rozwiązywania zagadnienia *estymacji pozy* na obrazach dwuwymiarowych [5, 6, 7, 40, 47, 134, 172], gdzie poprawia się ich jakość dzięki bogatszym deskryptorom do opisu wyglądu poszczególnych elementów [47], wprowadzeniu kontekstu w zależności od względnego położenia [172], a także zastosowaniu modelu kaskadowego *coarse-to-fine* [134]. Należy podkreślić, że podstawowym ograniczeniem *struktur obrazkowych* jest struktura drzewa, które nie pozwala zamodelować bardziej złożonych relacji. Niemniej rozważa się modele o bardziej zwartej sieci połączeń. Wtedy do szukania estymatora MAP stosuje się algorytmy optymalizacji dyskretnej, jak *metoda podziału i ograniczeń* (ang. branch and bound) [143, 149] czy heurystyki typu *A*-search* [16]. Dodatkowo ze względu na fakt, że wnioskowanie dla *struktur obrazkowych* wymaga przeszukiwania dyskretnej przestrzeni, nie jest możliwe efektywne przeniesienie tego modelu do problemu trójwymiarowej *estymacji pozy*. Stosuje się wtedy modele, gdzie położenie i orientacja poszczególnych części ciała są ciągłymi zmiennymi losowymi. W zależności od ich charakteru przyjmuje się różne strategie wnioskowania, m. in. algorytm *Nonparametric Belief Propagation* [142], próbkowanie losowe *Markov Chain Monte Carlo* (MCMC) [95, 96], zmodyfikowaną wersję metody *Expectation-Maximization* po wcześniejszej wokselizacji obrazów [34].

W porównaniu do metod opartych na *drzewie kinematycznym*, *modele oparte na częściach* dają zazwyczaj dużo lepsze wyniki w dwuwymiarowym statycznym zadaniu *estymacji pozy*. Niestety ich skuteczność nie przenosi się na rozważany w pracy problem trójwymiarowy, w szczególności w dynamicznym *śledzeniu ruchu*. Wydaje się, że przyszłość bezznacznikowych systemów śledzących będzie oparta na połączeniu obu tych idei.

1.3 Teza, cel i zakres pracy

Nawiązując do literatury przedmiotu, w pracy rozważane jest podejście *generujące* bazujące na reprezentacji konfiguracji przy pomocy *drzewa kinematycznego*, ze strategią wnioskowania polegającą na wyznaczeniu całego rozkładu a posteriori na przestrzeni konfiguracji. Wymaga ono określenia rozkładu a priori dla ułożenia ciała, *modelu wiarygodności* pozwalającego uwzględnić nową obserwację oraz algorytmu wnioskowania, który pozwoli

na efektywne oszacowanie rozkładu a posteriori.

Weźmy pod uwagę dwa spostrzeżenia. Po pierwsze dla ustalonych rodzajów ruchu (np. chodzenie, bieganie) stopnie swobody opisujące konfigurację charakteryzują się silnymi wzajemnymi zależnościami, co powoduje, że rzeczywiste trajektorie ruchu tworzą niskowymiarowe *rozmaitości* w przestrzeni możliwych ułożeń ciała. Po drugie wygląd poszczególnych elementów ciała (np. głowa, dłonie, stopy) jest unikatowy i niezmienny dla konkretnej postaci, powinien zatem być opisany w sposób, który pozwala na jednoznaczne rozpoznanie tych elementów. Prowadzi to do sformułowania następującej **tezy**:

> "*Zastosowanie wyspecjalizowanej wiedzy apriorycznej dotyczącej kinematyki lub wyglądu ciała człowieka pozwala na istotną poprawę jakości śledzenia ruchu w podejściu generującym.*"

W konsekwencji **celem rozprawy** jest opracowanie modeli i algorytmów, które pozwolą na wydobycie wiedzy apriorycznej dotyczącej kinematyki i wyglądu ciała człowieka na podstawie sekwencji treningowych oraz uwzględnienie jej w procesie *śledzenia ruchu*. W szczególności wyróżnić można następujące zadania badawcze:

1. Opracowanie algorytmu *śledzenia ruchu człowieka* uwzględniającego wiedzę o niskowymiarowej *rozmaitości*, na której rozkładają się rzeczywiste ułożenia ciała.
2. Opracowanie modeli pozwalających uwzględnić wiedzę o *rozmaitości* w rozkładzie a priori na przestrzeni konfiguracji.
3. Opracowanie *modelu wiarygodności* zawierającego informację o unikatowym wyglądzie poszczególnych części ciała.

W **zakres pracy** wchodzą następujące elementy:

1. Opracowanie probabilistycznego modelu zadającego łączny rozkład na trajektorię ruchu, sekwencję obserwacji i położenia na niskowymiarowej *rozmaitości*. Wyprowadzenie na jego podstawie zadania *filtrowania* (ang. filtering).
2. Opracowanie algorytmu bazującego na koncepcji *filtra cząsteczkowego*, rozwiązującego w sposób przybliżony zadanie *filtrowania* z uwzględnieniem niskowymiarowej *rozmaitości*.

3. Opracowanie *modeli dynamiki* w wysoko i niskowymiarowej przestrzeni, pozwalających na szacowanie rozkładu a priori w kolejnych taktach systemu oraz opracowanie metod ich uczenia.

4. Opracowanie *modelu wiarygodności* opartego na lokalnych deskryptorach części ciała. Ponadto opracowanie procedury dyskryminacyjnego uczenia w celu wyodrębnienia unikalnych cech poszczególnych elementów.

5. Przeprowadzenie badań empirycznych mających na celu zweryfikować jakość opracowanej metody śledzenia uwzględniającej niskowymiarową *rozmaitość* oraz metody śledzenia bazującej na zaproponowanym *modelu wiarygodności* i porównanie ich z metodami znanymi w literaturze.

Prezentowana praca poszerza aktualny stan wiedzy w zakresie *widzenia komputerowego* i *uczenia maszynowego*, w szczególności przedstawia metodę *filtra cząsteczkowego* uwzględniającego strukturę niskowymiarowej rozmaitości. Rezultaty w niej zawarte mogą być wykorzystane jako fragmenty systemu do bezznacznikowego *śledzenia ruchu człowieka*.

1.4 Układ pracy

Praca składa się z ośmiu rozdziałów. Kolejne rozdziały zawierają odpowiednio:

Rozdział 2. Opisana została koncepcja bezznacznikowego systemu śledzącego ruch człowieka. Przedstawiono charakterystykę danych na wejściu i wyjściu.

Rozdział 3. Sformułowano problemy *estymacji pozy* i *śledzenia ruchu człowieka* oraz oryginalny problem uwzględniający wiedzę aprioryczną o kinematyce ciała w postaci niskowymiarowej *rozmaitości*.

Rozdział 4. Przedstawiono trzy algorytmy śledzące. Dwa znane z literatury i jeden autorski rozwiązujący problem śledzenia ruchu uwzględniający strukturę *rozmaitości*.

Rozdział 5. Przestawiono trzy *modele wiarygodności*. Dwa znane z literatury i jeden autorski uwzględniający wiedzę aprioryczną o lokalnym wyglądzie wyróżnionych fragmentów ciała.

Rozdział 6. Przedstawiono dwa podejścia do modelowania dynamiki. Pierwsze standardowe podejście z literatury, a drugie autorskie uwzględniające dynamikę po niskowymiarowej *rozmaitości*.

Rozdział 7. Zaprezentowano wyniki badań empirycznych weryfikujących jakość zaproponowanego algorytmu śledzącego oraz zaproponowanego *modelu wiarygodności* w zadaniu *śledzenia ruchu człowieka*.

Rozdział 8. Podano uwagi końcowe ze wskazaniem oryginalnego wkładu pracy w dziedzinę oraz wskazano potencjalne kierunki dalszych badań.

Rozdział 2

Koncepcja systemu śledzącego

W tym rozdziale przedstawiona została koncepcja systemu śledzącego ruch człowieka jako systemu wejściowo-wyjściowego. W pierwszej części wprowadzono pojęcie *wektora stanu* człowieka jako zestawu wartości, które jest oczekiwane na wyjściu systemu. W drugiej części opisany został system pomiarowy, rejestrujący ruch człowieka w postaci obrazów wideo, które traktowane są jako wejście do systemu.

2.1 Reprezentacja ciała człowieka

Wektor stanu jest kluczowym pojęciem dla problemów *estymacji pozy* i *śledzenia ruchu*. Pod tym terminem rozumiemy zestaw wartości liczbowych, które wystarczają do zakodowania pełnej informacji o bieżącej konfiguracji ciała człowieka, zwanej inaczej stanem człowieka. Celem rozważanego w pracy problemu jest odtworzenie stanu na podstawie dostępnych obserwacji, dlatego *wektor stanu* będzie stanowił zestaw informacji, które chcemy uzyskać na wyjściu projektowanego systemu.

Wykorzystywane pojęcie *wektora stanu* jest terminem zaczerpniętym z automatyki i teorii sterowania [22, 25, 65], oznaczającym zestaw wielkości reprezentujących stan *systemu dynamicznego* (ang. dynamical system). Należy jednak podkreślić, że rozważane w pracy problemy nie są problemami sterowania. Dlatego w literaturze specyficznej dla zagadnień *estymacji pozy* i *śledzenia ruchu człowieka* często stosuje się pojęcie pozy (ang. pose), które w istocie oznacza *wektor stanu* opisujący bieżącą konfigurację ludzkiego ciała.

Wektor stanu powinien posiadać następujące dwie właściwości:

1. Być minimalną reprezentacją, tj. zawierać najmniejszy możliwy zbiór zmiennych potrzebny do pełnego opisania bieżącego stanu. Pozwala to na wnioskowanie o możliwie najmniejszej liczbie zmiennych i w konsekwencji prowadzi do modeli o prostszej postaci.

2. Posiadać niezależne składowe, tj. zmiana pojedynczej zmiennej w wektorze stanu nie powinna wymagać jednoczesnej zmiany innych zmiennych. Ma to znaczenie w procesie przeszukiwania przestrzeni stanów.

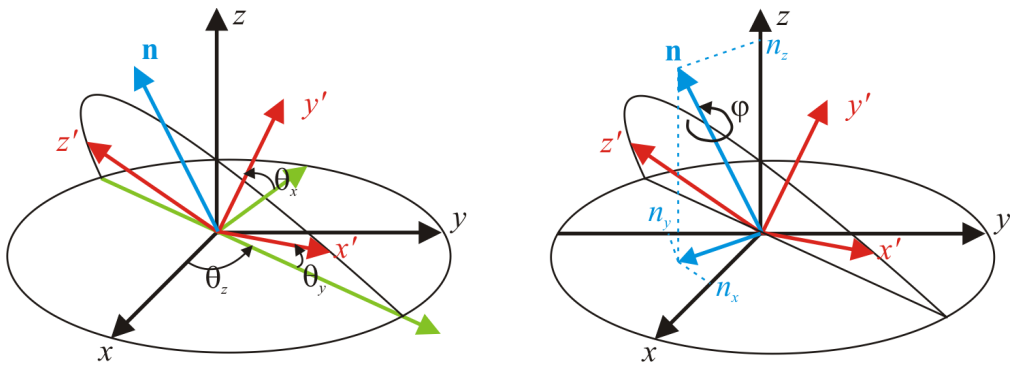
Dodatkowo pożądaną cechą opisu konfiguracji jest uniezależnienie go od wymiarów części ciała. Pozwala to na porównywanie ze sobą trajektorii ruchu różnych ludzi i ma kluczowe znaczenie na etapie uczenia modeli, jak również w późniejszych zastosowaniach.

2.1.1 Obroty w przestrzeni trójwymiarowej

Fundamentalnym problemem w konstruowaniu opisu konfiguracji ciała człowieka jest wybór reprezentacji obrotu w przestrzeni trójwymiarowej. W tym celu wprowadzamy dwa ortogonalne układy współrzędny: pierwszy (x, y, z), drugi (x', y', z'). Układy te mają wspólny środek i są względem siebie obrócone (rysunek 2.1). Przez reprezentację obrotu będziemy rozumieć sposób zakodowania ciągu operacji potrzebnych do przekształcenia pierwszego układu w drugi.

Pierwszym sposobem przedstawienia trójwymiarowego obrotu są *kąty Eulera* [62, 136]. Idea tej reprezentacji polega na wykonaniu serii trzech obrotów wokół ustalonych osi o zadane kąty. Ponieważ wybór osi, względem których wykonywane są kolejne obroty, nie jest jednoznaczny, dlatego istnieje wiele różnych reprezentacji obrotów przy pomocy *kątów Eulera*. Przykładowo konwencja $z - x - z$ oznacza, że pierwszy obrót wykonujemy wokół osi z, drugi wokół x, a trzeci ponownie wokół osi z. W pracy została przyjęta konwencja $z - x - y$.

Wielkości obrotów wokół osi x, y, z, wyrażamy wartościami liczbowymi $\theta_x, \theta_y, \theta_z$ z przedziału $[-\pi, \pi]$ (rysunek 2.1a). Reprezentacja przy pomocy *kątów Eulera* jest dzięki temu minimalna, tj. wystarczy podać trzy liczby $\theta_x, \theta_y, \theta_z$, aby wyrazić dowolny trójwymiarowy obrót. Wadą tej reprezentacji jest natomiast brak możliwości porównania dwóch obrotów



Rysunek 2.1: Reprezentacje obrotu w przestrzeni trójwymiarowej. (a) Kąty Eulera. (b) Kwaterniony.

przy pomocy metryki euklidesowej. Przykładowo weźmy dwa obroty wokół osi x, jeden o -0.9π, a drugi o 0.9π. Jeśli porównamy je przy pomocy metryki euklidesowej, to odległość między nimi wynosi 1.8π, mimo że rzeczywista odległość jest równa 0.2π. W przypadku obrotów trójwymiarowych sytuacja komplikuje się jeszcze bardziej. Istnieje możliwość zdefiniowania skomplikowanej metryki do porównania *kątów Eulera*, ale jej wyliczenie wiąże się z istotnym wzrostem złożoności obliczeniowej i może powodować brak wydajności algorytmów, które podczas swojego działania wymagają tysięcy porównań pomiędzy różnymi obrotami.

Można pokazać, że dowolny obrót w przestrzeni trójwymiarowej (zadany przykładowo przy pomocy *kątów Eulera*) jest równoważny obrotowi wokół pewnej osi wyznaczonej przez jednostkowy wektor $\mathbf{n} = (n_x, n_y, n_z)$ o pewien kąt φ (rysunek 2.1b). Ten rezultat znany jest jako *twierdzenie Eulera* (ang. Euler's rotation theorem).

Prowadzi to do alternatywnej reprezentacji obrotów przy pomocy jednostkowych *kwaternionów* [89, 136], które są wyrażone jako czwórka $\mathbf{q} = (q_w, q_x, q_y, q_z)$, gdzie odpowiednie

składowe mają interpretację:

$$q_w = \cos(\varphi/2), \tag{2.1}$$

$$q_x = n_x \sin(\varphi/2), \tag{2.2}$$

$$q_y = n_y \sin(\varphi/2), \tag{2.3}$$

$$q_z = n_z \sin(\varphi/2). \tag{2.4}$$

Naturalną metryką do porównywania *kwaternionów* jest metryka euklidesowa. Dzięki temu możliwe jest efektywne porównywanie obrotów dla tej reprezentacji. Wadą *kwaternionów* jest fakt, że nie stanowią minimalnej reprezentacji obrotu, tj. jednostkowy *kwaternion* kodowany jest przy pomocy czterech liczb, a do wyrażenia dowolnego obrotu wystarczają trzy wartości. Dodatkowo jednostkowy *kwaternion* jest unormowany:

$$\|\mathbf{q}\| = \sqrt{q_w^2 + q_x^2 + q_y^2 + q_z^2} = 1, \tag{2.5}$$

co powoduje, że nie można traktować odpowiednich składowych *kwaternionu* niezależnie, np. zaburzając je addytywnym szumem gaussowskim. Problem ten nie występuje w przypadku kątów Eulera.

W przypadku, gdy obrót wykonywany jest o kąt $0 < \varphi < \pi$, wtedy $q_w > 0$ i możemy zastosować następującą aproksymację dla *kwaternionu*:

$$\overline{\mathbf{q}} = \left(\frac{q_x}{q_w}, \frac{q_y}{q_w}, \frac{q_z}{q_w} \right). \tag{2.6}$$

W tym przypadku dowolny obrót reprezentowany za pomocą zaproponowanej aproksymacji może być jednoznacznie odtworzony do jednostkowego *kwaternionu*. Zaletą tego przybliżenia jest to, że składowe $\overline{\mathbf{q}}$ mogą być odtąd traktowane niezależnie. Ponadto obroty nadal mogą być porównywane przy pomocy metryki euklidesowej, gdyż zaproponowana postać stanowi jedynie przeskalowanie składowych osi obrotu. Podobne aproksymacje zostały zastosowane w pracach [41, 141].

Własnością charakterystyczną dla *kwaternionów* jest to, że stanowią one uogólnienie liczb zespolonych i mogą być przedstawione w następującej postaci:

$$\mathbf{q} = q_w + q_x i + q_y j + q_z k, \tag{2.7}$$

gdzie i, j, k oznaczają odpowiednio trzy niezależne jednostki urojone. Podobnie, jak w przypadku liczb zespolonych jednostki urojone posiadają własność $i^2 = j^2 = k^2 = -1$, a także spełniają inne wzajemne relacje [89]. Ta reprezentacja pozwala na bardzo efektywne składanie obrotów w przestrzeni trójwymiarowej poprzez mnożenie *kwaternionów* w postaci (2.7) (podobnie, jak ma to miejsce w przypadku składania obrotów na płaszczyźnie poprzez mnożenie liczb zespolonych). Złożenie dwóch obrotów wymaga wykonania jedynie 28 elementarnych operacji algebraicznych, jak dodawanie i mnożenie.

Ważnym problemem jest również obracanie ustalonego punktu $\mathbf{v} = (v_x, v_y, v_z)^{\mathrm{T}}$, tj. przekształcenie jego składowych wyrażonych w układzie (x, y, z) do składowych w innym układzie (x', y', z'). Ma to szczególne znaczenie w metodach grafiki trójwymiarowej, kiedy chcemy wizualizować pewien obiekt widziany z danej perspektywy. W tym przypadku użycie *kwaternionów* wymaga wykonania aż 30 elementarnych operacji algebraicznych i nie jest najbardziej efektywną metodą.

Problem ten prowadzi do ostatniej reprezentacji obrotów przy pomocy macierzy rotacji [8, 136]:

$$\mathbf{R} = [\mathbf{r}_x \ \mathbf{r}_y \ \mathbf{r}_z], \tag{2.8}$$

gdzie kolumny $\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z$ oznaczają odpowiednio końce wersorów rozpinających układ (x, y, z) wyrażone w układzie (x', y', z'). W ten sposób punkt $\mathbf{v}$ wyrażony w układzie (x, y, z) może być obrócony do układu (x', y', z') poprzez przemnożenie go przez macierz obrotu: $\mathbf{v}' = \mathbf{R}\mathbf{v}$. Wymaga to wykonania jedynie 15 elementarnych operacji algebraicznych i przez to jest powszechnie stosowaną reprezentacją do obracania punktów.

Dodatkowo ze względu na interpretację kolumn macierzy rotacji, ta reprezentacja jest najłatwiejsza do otrzymania, wykorzystując pomiary z *systemów MOCAP*, które zazwyczaj podają położenia w przestrzeni czterech punktów (początek układu współrzędnych i po jednym punkcie na każdej z osi).

Podobnie, jak w przypadku *kwaternionów*, składanie obrotów wyrażonych przy pomocy macierzy rotacji polega na przemnożeniu macierzy przez siebie. Dodatkowo macierz obrotu jest macierzą ortogonalną, tj. $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$, czyli macierz do niej odwrotna (odpowiadająca odwrotnemu obrotowi) wyrażona jest poprzez jej transpozycję. Ma to znaczenie w przypadku, gdy macierze rotacji $\mathbf{R}$ i $\mathbf{R}'$ odpowiadające układom (x, y, z) i (x', y', z') wyrażone są względem standardowego układu współrzędnych rozpiętego przez wersory $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$.

Wtedy rotację z układu (x, y, z) do (x', y', z') uzyskujemy poprzez złożenie $\mathbf{R}'^{\mathrm{T}}\mathbf{R}$.

Dowolny obrót może być wyrażony poprzez każdą z trzech wymienionych reprezentacji. Z faktu, że różnią się one własnościami, a w konsekwencji także zastosowaniami, istotnym problemem jest posiadanie zestawu transformacji, które pozwalają na przechodzenie pomiędzy reprezentacjami. Ponieważ w pracy stosowana jest konwencja $z - x - y$, dlatego transformacje będą dostosowane do niej.

1. Przejście z reprezentacji przy pomocy *kątów Eulera* $\theta_x, \theta_y, \theta_z$ do macierzy rotacji $\mathbf{R}$ odbywa się poprzez zastosowanie macierzy rotacji wokół ustalonych osi $\mathbf{R}_x$, $\mathbf{R}_y$, $\mathbf{R}_z$, gdzie odpowiednio:

$$\mathbf{R}_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix}, \tag{2.9}$$

$$\mathbf{R}_y = \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix}, \tag{2.10}$$

$$\mathbf{R}_z = \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.11}$$

Wtedy końcową rotację składamy z trzech składowych zgodnie z konwencją $z-x-y$:

$$\mathbf{R} = \mathbf{R}_z\mathbf{R}_x\mathbf{R}_y. \tag{2.12}$$

Przejście w odwrotną stronę uzyskujemy poprzez zastosowanie następujący wyrażeń bezpośrednio do macierzy rotacji $\mathbf{R} = [r_{ij}]$:

$$\theta_x = \arcsin(-r_{32}), \tag{2.13}$$

$$\theta_y = \arctan 2(r_{12}, r_{22}), \tag{2.14}$$

$$\theta_z = \arctan 2(r_{31}, r_{33}), \tag{2.15}$$

gdzie funkcja $\arctan 2(x, y)$ zdefiniowana jest następująco:

$$\arctan 2(x,y) = \begin{cases} \arctan\left(\frac{x}{y}\right), & y > 0 \\ \arctan\left(\frac{x}{y}\right) + \pi, & x \geqslant 0, y < 0 \\ \arctan\left(\frac{x}{y}\right) - \pi, & x < 0, y < 0 \\ +\frac{\pi}{2}, & x > 0, y = 0 \\ -\frac{\pi}{2}, & x < 0, y = 0 \\ \text{nieokreślona}, & x = 0, y = 0 \end{cases} \tag{2.16}$$

Warto zwrócić uwagę, że o *kątach Eulera* zakładamy, że należą do przedziału $[-\pi, \pi]$. Z tego powodu równanie (2.13) będzie miało dwa rozwiązania, ponieważ funkcja $\arcsin$ standardowo określona jest na przedziale $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. W wielu praktycznych problemach obroty są wykonywane o niewielkie kąty (również w problemie *śledzenia ruchu człowieka*) i wtedy rozważa się jedynie rozwiązanie w przedziale $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

2. Do zdefiniowania przejścia pomiędzy reprezentacją przy pomocy *kwaternionów* $\mathbf{q} = (q_w, q_x, q_y, q_z)$ i macierzy rotacji wprowadźmy następujące oznaczenia:

$$\breve{\mathbf{q}} = (q_x, q_y, q_z)^{\mathrm{T}}, \tag{2.17}$$

$$\mathbf{Q} = \begin{bmatrix} 0 & -q_z & q_y \\ q_z & 0 & -q_x \\ -q_y & q_x & 0 \end{bmatrix}. \tag{2.18}$$

Wtedy macierz rotacji może być wyrażona przy pomocy następującej formuły:

$$\mathbf{R} = (q_w^2 - \breve{\mathbf{q}}^{\mathrm{T}}\breve{\mathbf{q}})\mathbf{I}_{3\times 3} + 2\breve{\mathbf{q}}\breve{\mathbf{q}}^{\mathrm{T}} + 2q_w\mathbf{Q}. \tag{2.19}$$

Powyższa zależność stanowi przekształconą wersję formuły Rodriguesa (ang. Rodrigues' rotation formula), która określa związek pomiędzy macierzą rotacji oraz wektorem $\mathbf{n}$ i kątem φ określonymi przez *twierdzenie Eulera*.

Ponieważ $\mathbf{q}$ i $-\mathbf{q}$ reprezentują ten sam obrót, dlatego przejście w odwrotną stronę, tj. od macierzy rotacji do *kwaternionu*, nie jest jednoznaczne. Zakładamy, że $q_w > 0$,

wtedy reprezentację kwaternionową możemy otrzymać z następujących zależności:

$$q_w = \frac{1}{2}\sqrt{1 + r_{11} + r_{22} + r_{33}}, \tag{2.20}$$

$$q_x = \frac{1}{4q_w}(r_{32} - r_{23}), \tag{2.21}$$

$$q_y = \frac{1}{4q_w}(r_{13} - r_{31}), \tag{2.22}$$

$$q_z = \frac{1}{4q_w}(r_{21} - r_{12}). \tag{2.23}$$

3. Przejście od reprezentacji przy pomocy *kątów Eulera* do reprezentacji przy pomocy *kwaternionów* wymaga założenia o kolejności wykonywania obrotów, która została ustalona na $z - x - y$. Wtedy składowe *kwaternionu* otrzymujemy z następujących zależności:

$$q_w = \cos\left(\frac{\theta_x}{2}\right)\cos\left(\frac{\theta_y}{2}\right)\cos\left(\frac{\theta_z}{2}\right) - \sin\left(\frac{\theta_x}{2}\right)\sin\left(\frac{\theta_y}{2}\right)\sin\left(\frac{\theta_z}{2}\right), \tag{2.24}$$

$$q_x = \sin\left(\frac{\theta_x}{2}\right)\cos\left(\frac{\theta_y}{2}\right)\cos\left(\frac{\theta_z}{2}\right) - \cos\left(\frac{\theta_x}{2}\right)\sin\left(\frac{\theta_y}{2}\right)\sin\left(\frac{\theta_z}{2}\right), \tag{2.25}$$

$$q_y = \cos\left(\frac{\theta_x}{2}\right)\sin\left(\frac{\theta_y}{2}\right)\cos\left(\frac{\theta_z}{2}\right) + \sin\left(\frac{\theta_x}{2}\right)\cos\left(\frac{\theta_y}{2}\right)\sin\left(\frac{\theta_z}{2}\right), \tag{2.26}$$

$$q_z = \cos\left(\frac{\theta_x}{2}\right)\cos\left(\frac{\theta_y}{2}\right)\sin\left(\frac{\theta_z}{2}\right) + \sin\left(\frac{\theta_x}{2}\right)\sin\left(\frac{\theta_y}{2}\right)\cos\left(\frac{\theta_z}{2}\right). \tag{2.27}$$

Analogicznie możemy wykonać przejście od postaci kwaternionowej do *kątów Eulera* przy pomocy następujących wyrażeń:

$$\theta_x = \arcsin\Big(2(q_w q_x + q_y q_z)\Big), \tag{2.28}$$

$$\theta_y = \arctan 2\Big(2(q_w q_y - q_x q_z), 1 - 2(q_x^2 + q_y^2)\Big), \tag{2.29}$$

$$\theta_z = \arctan 2\Big(2(q_w q_z - q_x q_y), 1 - 2(q_x^2 + q_z^2)\Big), \tag{2.30}$$

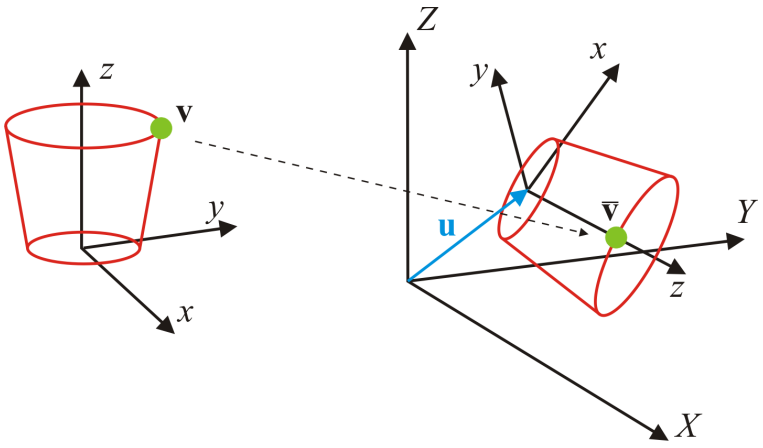
gdzie funkcja $\arctan 2$ została zdefiniowana przy pomocy zależności (2.16).

Podsumowując, każda z trzech przedstawionych reprezentacji obrotów ma swoje wady i zalety. *Kąty Eulera* stanowią minimalną reprezentację obrotu, a ich składowe są od siebie

niezależne, nie mogą być jednak porównywane przy pomocy metryki euklidesowej. Jednostkowe *kwaterniony* mogą być porównywane przy pomocy tej metryki, natomiast nie stanowią minimalnej reprezentacji i ich składowe są od siebie zależne poprzez warunek (2.5). W celu wyeliminowania tych problemów można zastosować aproksymację (2.6), jednakże ogranicza ona zakres możliwych obrotów. Do efektywnego obracania punktów najlepiej wykorzystać reprezentację przy pomocy macierzy obrotu, gdyż wymaga ona wykonania najmniejszej liczby elementarnych działań algebraicznych w porównaniu do pozostałych reprezentacji.

2.1.2 Drzewo kinematyczne

Pojedynczą część ciała będziemy modelowali przy pomocy elementu sztywnego (ang. rigid element). Pojęcie to odnosi się do obiektów fizycznych o dodatniej objętości, których fragmenty nie przemieszczają się względem siebie. Stanowi to uproszczenie rzeczywistego problemu, gdzie podczas ruchu występują drobne odkształcenia w obrębie pojedynczych kończyn. Efekt ten jest szczególnie widoczny, jeśli człowiek jest ubrany w strój, który ściśle nie przylega do ciała.



Rysunek 2.2: Transformacja elementu sztywnego z lokalnego do globalnego układu współrzędnych.

Ponieważ fragmenty elementu sztywnego nie przemieszczają się względem siebie, to możemy na stałe zorientować na nim pewien ortogonalny układ współrzędnych (x, y, z),

zwany dalej lokalnym układem współrzędnych, względem którego wszystkie fragmenty będą nieruchome. Wtedy dowolny punkt z tego elementu może być opisany przy pomocy współrzędnych z lokalnego układu, tj. $\mathbf{v} = (v_x, v_y, v_z)^{\mathrm{T}}$. W konsekwencji możemy zdefiniować element sztywny jako zbiór punktów w lokalnym układzie współrzędnych $\mathcal{V}$.

Załóżmy teraz, że element sztywny wykonuje ruch w pewnej ograniczonej trójwymiarowej przestrzeni, zwanej dalej sceną, z którą skojarzony będzie na stałe globalny układ współrzędnych (X, Y, Z). Wtedy dowolny punkt $\mathbf{v} \in \mathcal{V}$ może być wyrażony w globalnym układzie współrzędnych poprzez zależność:

$$\overline{\mathbf{v}} = \mathbf{R}\mathbf{v} + \mathbf{u}, \tag{2.31}$$

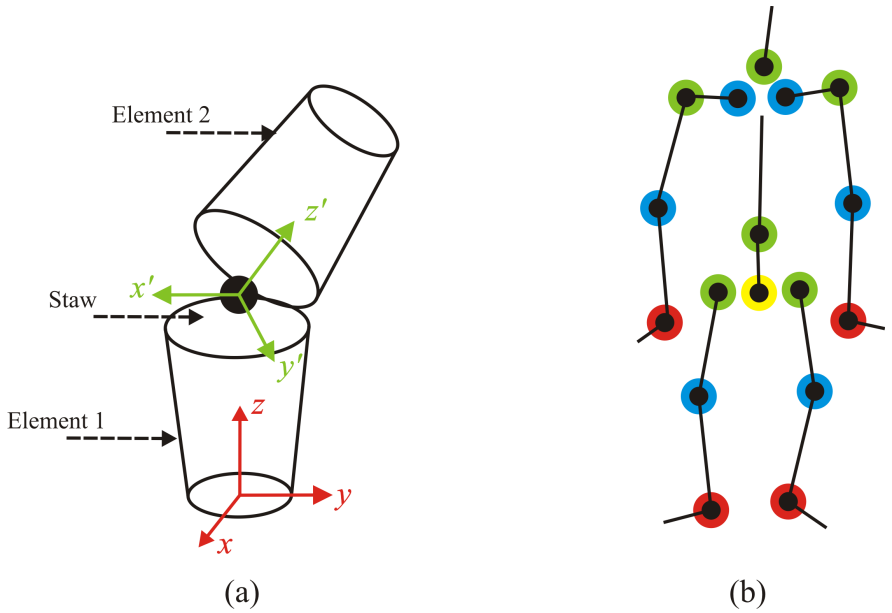
gdzie $\mathbf{R}$ oznacza macierz rotacji pomiędzy układami, a $\mathbf{u}$ oznacza wektor przesunięcia (rysunek 2.2). Możemy w ten sposób zdefiniować zbiór wszystkich punktów $\overline{\mathcal{V}}$ z elementu $\mathcal{V}$ wyrażonych w globalnym układzie współrzędnych.

Ponieważ fragmenty elementu sztywnego nie przemieszczają się między sobą, zatem ruch w globalnym układzie współrzędnych może być zdefiniowany poprzez zmianę w czasie macierzy rotacji $\mathbf{R}$ oraz wektora przesunięcia $\mathbf{u}$. Wynika stąd wniosek, że ruch może być określony poprzez zmianę sześciu stopni swobody, tj. trzy składowe wektora przesunięcia, $\mathbf{u} \in \mathbb{R}^3$, oraz trzy stopnie swobody macierzy rotacji, $\mathbf{R} \in SO(3)$, które wynikają natychmiast z dekompozycji (2.12).

Rozważmy teraz układ dwóch elementów sztywnych połączonych ze sobą przy pomocy ruchomego stawu (rysunek 2.3a). Taki układ określany jest często obiektem przegubowo połączonym (ang. articulated object). Niech $\mathbf{R}_1$, $\mathbf{R}_2$ oraz $\mathbf{u}_1$, $\mathbf{u}_2$ oznaczają odpowiednio macierze rotacji i wektory przesunięcia pozwalające na przejście od lokalnych układów współrzędnych do globalnego układu współrzędnych odpowiednio dla elementów 1 i 2 (rysunek 2.3a). Przejście to uzyskujemy poprzez zastosowanie zależności (2.31).

Powyższa reprezentacja ma 12 parametrów (6 dla stworzenia macierzy rotacji i 6 dla wektorów przesunięć). Zauważmy, że położenie stawu (gdzie zaczepiony jest lokalny układ dla elementu 2) pozostaje niezmienne w lokalnym układzie dla elementu 1, co sugeruje, że rzeczywista liczba stopni swobody jest mniejsza niż 12.

Załóżmy teraz, że element 1 jest nadrzędny w stosunku do elementu 2. Będziemy określali, że element 1 jest rodzicem dla elementu 2. Niech $\mathbf{v}_2$ oznacza położenie punktu w lokalnym układzie dla elementu 2. Wtedy położenie tego punktu, $\mathbf{v}_1$, w lokalnym układzie



Rysunek 2.3: Obiekty przegubowo połączone. (a) Pojedyczny staw. (b) Drzewo kinematyczne dla człowieka.

dla elementu 1 uzyskujemy poprzez dwukrotne zastosowanie wzoru (2.31):

$$\begin{aligned}\mathbf{v}_1 &= \mathbf{R}_1^{\mathrm{T}}(\mathbf{R}_2\mathbf{v}_2 + \mathbf{u}_2 - \mathbf{u}_1)\\ &= \mathbf{R}_1^{\mathrm{T}}\mathbf{R}_2\mathbf{v}_2 + \mathbf{R}_1^{\mathrm{T}}(\mathbf{u}_2 - \mathbf{u}_1)\\ &= \mathbf{R}_{2,1}\mathbf{v}_2 + \mathbf{u}_{2,1}.\end{aligned} \tag{2.32}$$

W powyższym równaniu $\mathbf{R}_{2,1}$ oznacza macierz rotacji z lokalnego układu dla elementu 2 do lokalnego układu dla elementu 1. Natomiast $\mathbf{u}_{2,1}$ oznacza położenie początku układu dla elementu 2 w układzie dla elementu 1 (położenie stawu), które jest stałe.

Zatem do wyznaczenia położenia punktu z elementu 2 w globalnym układzie współrzędnych należy zastosować wzór (2.32), a następnie (2.31). W konsekwencji do opisanie ruchu dwóch elementów sztywnych połączonych przegubowo w globalnym układzie współrzędnych potrzebujemy jedynie 9 stopni swobody (po 3 stopnie swobody do opisania odpowiednio $\mathbf{R}_1$, $\mathbf{u}_1$, $\mathbf{R}_{2,1}$).

Zauważmy, że macierz $\mathbf{R}_{2,1}$ zadaje charakterystykę dla stawu, tj. podaje w jaki sposób staw jest zgięty. Należy zwrócić uwagę, że w rzeczywistości stawy mogą mieć jeden (ang.

revolute joint), dwa (ang. hardy-spicer joint) lub trzy (ang. spherical joint) stopnie swobody [34, 136]. W konsekwencji macierz $\mathbf{R}_{2,1}$ może być zadana jako złożenie obrotów w jednej, dwóch lub trzech płaszczyznach i posiadać od jednego do trzech stopni swobody.

Ciało człowieka będziemy modelować jako zbiór połączonych przegubowo elementów sztywnych. Taki układ nazywamy *drzewem kinematycznym* (ang. kinematic tree) i jest to najpowszechniej stosowany sposób reprezentowania ciała [19, 24, 26, 44, 58, 63, 66, 71, 72, 77, 84, 86, 94, 97, 114, 128, 139, 140, 151, 153, 161, 177].

Na rysunku 2.3b zostało przedstawione *drzewo kinematyczne* stosowane w pracy. Wyróżnione zostały następujące elementy: miednica, tułów, głowa, barki, ramiona, przedramiona, ręce, uda, łydki i stopy. Każdy z nich połączony jest ze swoim rodzicem przy pomocy ruchomego stawu. Kolorem zielonym zostały wyróżnione stawy o trzech stopniach swobody, niebieski o dwóch stopniach swobody, a czerwonym o jednym stopniu swobody. Pierwszym elementem w *drzewie kinematycznym* jest miednica. Kolorem żółtym został oznaczony „pseudostaw", który reprezentuje rotację względem globalnego układu współrzędnych.

Rozważmy teraz punkt $\mathbf{v}_i$ należący do i-tego elementu sztywnego, wtedy współrzędne tego punktu w układzie współrzędnych jego rodzica wyrażamy przy pomocy zależności:

$$\mathbf{v}_{\mathrm{pa}(i)} = \mathbf{R}_{i,\mathrm{pa}(i)}\mathbf{v}_i + \mathbf{u}_{i,\mathrm{pa}(i)}, \tag{2.33}$$

gdzie $\mathrm{pa}(i)$ oznacza indeks rodzica i-tego elementu. Macierz rotacji $\mathbf{R}_{i,\mathrm{pa}(i)}$ odpowiada za zgięcie stawu łączącego i-ty element z jego rodzicem. Zatem, aby wyrazić współrzędne punktu $\mathbf{v}_i$ w globalnym układzie współrzędnych należy rekurencyjnie zastosować zależność (2.33), aż osiągniemy współrzędne punktu w lokalnym układzie dla pierwszego elementu w drzewie, a następnie skorzystać ze wzoru (2.31). Stosując tę procedurę do wszystkich punktów z elementu $\mathcal{V}_i$, otrzymamy reprezentację elementu w globalnym układzie współrzędnych $\overline{\mathcal{V}}_i$.

W konsekwencji, aby opisać konfigurację człowieka na obserwowanej scenie, wystarczy, że znamy położenie $\mathbf{u}_0$ i rotację $\mathbf{R}_0$ pierwszego elementu w *drzewie kinematycznym* oraz względne rotacje $\mathbf{R}_{1,\mathrm{pi}(1)}, \ldots, \mathbf{R}_{K,\mathrm{pa}(K)}$ dla pozostałych K elementów.

Dodatkowo musimy znać $\mathbf{u}_{i,\mathrm{pa}(i)}$, tj. położenia początków lokalnych układów współrzędnych w układach współrzędnych ich rodziców. W pracy zakładamy, że elementy zaczynają się w początku swojego układu i są skierowane wzdłuż osi z, wtedy możemy przyjąć, że:

$$\mathbf{u}_{i,\mathrm{pa}(i)} = (0, 0, h_{pa(i)})^{\mathrm{T}}, \tag{2.34}$$

gdzie h_i oznacza długość i-tej kończyny, o której zakładamy, że jest znana.

2.1.3 Wektor stanu

Korzystając z reprezentacji ciała człowieka przy pomocy *drzewa kinematycznego* możemy sformułować *wektor stanu*, który opisuje bieżącą konfigurację.

Przez $\boldsymbol{\theta}_i = (\theta_{i,x}, \theta_{i,y}, \theta_{i,z})$ oznaczmy *kąty Eulera* odpowiadające macierzom $\mathbf{R}_{i,\mathrm{pa(i)}}$ zgodnie z zależnością (2.12). Definiujemy zredukowany *wektor stanu*:

$$\check{\mathbf{x}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^{\mathrm{T}}. \tag{2.35}$$

Zauważmy, że reprezentuje on jedynie wewnętrzną konfigurację ciała, tj. niezależną od przesunięcia i rotacji względem globalnego układu współrzędnych. Jest to minimalny zbiór niezależnych wartości potrzebnych do jej opisu.

Dodatkowo przez $\boldsymbol{\theta}_0 = (\theta_{0,x}, \theta_{0,y}, \theta_{0,z})$ oznaczmy *kąty Eulera*, odpowiadające macierzy $\mathbf{R}_0$. Wtedy $\mathbf{u}_0$ i $\boldsymbol{\theta}_0$ stanowią minimalny zbiór zmiennych potrzebny do opisania położenia i rotacji początku *drzewa kinematycznego* względem globalnego układu współrzędnych.

Wektor stanu definiujemy w następujący sposób:

$$\mathbf{x} = (\mathbf{u}_0, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^{\mathrm{T}}. \tag{2.36}$$

Ten opis uwzględnia dodatkowo przesunięcie i rotację względem globalnego układu współrzędnych, a więc zawiera pełną informację potrzebną do opisu konfiguracji ciała człowieka na scenie. Dodatkowo jest to reprezentacja minimalna oraz składowe są od siebie niezależne.

Alternatywnie, możemy zastąpić reprezentację przy pomocy *kątów Eulera* na zredukowane *kwaterniony* zdefiniowane przez (2.6). Wtedy zredukowany *wektor stanu* i pełny *wektor stanu* mają postać:

$$\check{\mathbf{x}} = (\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_K)^{\mathrm{T}}, \tag{2.37}$$

$$\mathbf{x} = (\mathbf{u}_0, \overline{\mathbf{q}}_0, \overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_K)^{\mathrm{T}}. \tag{2.38}$$

Należy zauważać, że te reprezentacje są minimalne jedynie wtedy, gdy dla wszystkich stawów założymy jeden lub trzy stopnie swobody. W przypadku stawów o dwóch stopniach

swobody, ich rotacja musi być reprezentowana przez trzy składowe zredukowanego *kwaternionu*, a przez to nie jest reprezentacją minimalną. Zastosowanie zredukowanych *kwaternionów* powoduje, że podobnie jak w przypadku *kątów Eulera*, składowe są od siebie niezależne.

W pracy zostało użyte to samo oznaczenie x na *wektor stanu* w postaci (2.36) i (2.38). Wynika to z faktu, że większość dalszych rozważań jest niezależna od reprezentacji obrotu. W miejscach, gdzie ma to znaczenie, zostało zaznaczone, z której reprezentacji korzystamy. Analogiczna sytuacja dotyczy zredukowanego *wektora stanu*.

Na koniec należy zwrócić uwagę na wymiar przestrzeni *wektora stanu*. W zależności od przyjętej reprezentacji obrotów, a także od stopni swobody uwzględnionych w poszczególnych stawach, wymiar waha się w granicach 40-60 zmiennych. Wysoka wymiarowość jest jedną z podstawowych przyczyn, która stanowi o trudności problemów *estymacji pozy* i *śledzenia ruchu człowieka*. Pojawiają się tutaj problemy, które często określa się wspólną nazwą *klątwy wymiarowości* (ang. curse of dimensionality) [17].

2.2 System pomiarowy

Odtwarzanie konfiguracji ciała odbywa się na podstawie obrazów z kilku zsynchronizowanych kamer, obserwujących człowieka z różnych perspektyw. Kamery powinny być rozstawione w taki sposób, aby każda z nich wnosiła możliwie najwięcej informacji, tj. obserwowały różne fragmenty sceny.

Zakładamy, że obraz z pojedynczej kamery nie jest w żaden sposób przetwarzany i na wejściu systemu jest podawany w formacie RGB lub w odcieniach szarości. Są to najpopularniejsze formy cyfrowego zapisu (bez kompresji) odpowiednio obrazu kolorowego i czarno białego [55, 138].

Formalnie kolorowy obraz możemy zapisać jako trzywymiarowy tensor zawierający wartości poszczególnych pikseli:

$$\mathrm{I} = [\mathrm{I}^c_{ij}], \tag{2.39}$$

gdzie i, j określają położenie piksela, natomiast $c \in \{R, G, B\}$ określa kolor. W przypadku obrazu czarno białego występuje tylko jeden kolor kodujący odcień szarości.

Przez $\mathcal{I}$ oznaczmy zbiór wszystkich dostępnych obrazów z kamer w danej chwili, gdzie

każdy z nich jest w postaci (2.39). Innymi słowy, $\mathcal{I}$ jest pojedynczym pomiarem stanowiącym wejście dla tworzonego systemu, na podstawie którego jest szacowany bieżący wektor stanu $\mathbf{x}$.

Należy zwrócić uwagę, że rozważany problem jest źle uwarunkowany (ang. ill-posed problem). Wynika to z faktu, że chcemy wnioskować o trójwymiarowej rzeczywistości, bazując na dwuwymiarowych obrazach. Oznacza to, że istnieje nieskończenie wiele konfiguracji $\mathbf{x}$, którym odpowiada ta sama obserwacja $\mathcal{I}$. Oczywiście im większą liczbą kamer dysponujemy, tym ta niejednoznaczność jest mniejsza.

Jednym z możliwych podejść do rozwiązania powyższych trudności jest odwrócenie problemu, tj. próba wygenerowania obrazu, który powinien pojawiać się na ustalonej kamerze, jeśli założymy, że znamy prawidłową konfigurację $\mathbf{x}$. Stosowanie tego podejścia wymaga następujących ustaleń:

1. Dobrania postaci poszczególnych elementów $\mathcal{V}$, które z uwzględnieniem struktury drzewa kinematycznego składają się na *model ciała* (ang. body model).

2. Opisania wyglądu poszczególnych elementów, jak kolor, faktura, drobne szczegóły itp. Sposób opisu będzie określony przez *model wyglądu* (ang. appearance model).

3. Określenia położenia poszczególnych fragmentów ciała w lokalnym układzie współrzędnych dla ustalonej kamery. Wymaga to rozwiązania problemu *kalibracji kamer* (ang. camera calibration).

4. Zrzutowania trójwymiarowej rzeczywistości na dwuwymiarowy obraz przy użyciu projekcji perspektywicznej (ang. perspective projection), z uwzględnieniem właściwości charakterystycznych dla danej kamery (przeskalowanie, zniekształcenia).

W bieżącym podrozdziale został omówiony problem *kalibracji kamer* i projekcja perspektywiczna. Dobór *modelu ciała* i *modelu wyglądu* nie są związane z charakterystyką układu pomiarowego i zostały opisane w rozdziale 5.

2.2.1 Kalibracja kamer

Kalibracja kamer (ang. camera calibration) jest jednym z fundamentalnych problemów leżących u podstaw *widzenia komputerowego* (ang. computer vision) [55, 67, 138]. Dziedzi-

na ta nieformalnie mówiąc zajmuje się wnioskowaniem o trójwymiarowej rzeczywistości w oparciu o dwuwymiarowe obrazy. W tych zagadnieniach często konieczne jest posiadanie informacji, który fragment sceny jest widoczny z danej kamery. Zakładając, że z obserwowaną rzeczywistością związany jest globalny układ współrzędnych, interesuje nas odnalezienie relacji pomiędzy tym układem, a lokalnymi układami poszczególnych kamer.

W literaturze zostało zaproponowane wiele metod rozwiązania tego problemu. Wyróżnić wśród nich można trzy grupy algorytmów. W pierwszej grupie wykorzystuje się obiekty kalibracyjne, na których ustala się pewne punkty i bada relacje pomiędzy tymi punktami obserwowanymi na obrazach z różnych kamer. Proste metody używają trójwymiarowe obiekty kalibracyjne [159], bardziej zaawansowane wykorzystują szablony kalibracyjne dwuwymiarowe [175] i jednowymiarowe [176]. Drugą grupę metod stanowią techniki autokalibracji (ang. self-calibration), które bazują jedynie na obrazach z kamery, wykorzystując ograniczenia i zależności wynikające z geometrii projekcyjnej (ang. projective geometry) [101, 123]. Ostatnia klasa metod łączy ze sobą techniki pierwszej i drugiej grupy. Wyróżnić tu można m.in. metody stosujące znikające punkty (ang. vanishing points) [29, 98], kalibrujące na podstawie ludzkiego ruchu [102].

W pracy stosujemy technikę kalibracji z dwuwymiarowym szablonem kalibracyjnym [175]. Załóżmy, że $\mathbf{v}^{\mathrm{I}}$ oznacza współrzędne punktu w lokalnym układzie współrzędnych kamery, natomiast $\overline{\mathbf{v}}$ współrzędne w globalnym układzie związanym ze sceną. Dodatkowo wprowadźmy oznaczenie $\tilde{\mathbf{v}}^{\mathrm{I}} = (\tilde{v}_x^{\mathrm{I}}, \tilde{v}_y^{\mathrm{I}}, 1)^{\mathrm{T}}$, gdzie $\tilde{v}_x^{\mathrm{I}}$, $\tilde{v}_y^{\mathrm{I}}$ oznaczają współrzędne punkt na obrazie I. Wtedy zachodzi następującą zależność:

$$s\tilde{\mathbf{v}}^{\mathrm{I}} = \mathbf{A}\mathbf{v}^{\mathrm{I}}, \tag{2.40}$$

gdzie s jest ustalonym parametrem skali związanym z projekcją punktu na obraz. Macierz $\mathbf{A}$ jest tzw. macierzą parametrów wewnętrznych kamery (ang. camera intrinsic matrix) i ma postać:

$$\mathbf{A} = \begin{bmatrix} \alpha_c & \gamma_c & a_c \\ 0 & \beta_c & b_c \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.41}$$

W powyższej zależności (a_c, b_c) oznacza odpowiednio położenie początku lokalnego układu na obrazie I, α_c i β_c są współczynnikami skalującymi osie, a γ_c jest parametrem skośności.

Korzystając z faktu, że relacja między $\overline{\mathbf{v}}$ i $\mathbf{v}^{\mathrm{I}}$ opisana jest zależnością (2.31) oraz używając rozszerzoną reprezentację $\tilde{\mathbf{v}} = (\overline{v}_x, \overline{v}_y, \overline{v}_z, 1)^{\mathrm{T}}$ dla punktu $\overline{\mathbf{v}}$, otrzymujemy następującą zależność:

$$s\tilde{\mathbf{v}}^{\mathrm{I}} = \mathbf{A}[\mathbf{R}_{\mathrm{I}} \;\; \mathbf{u}_{\mathrm{I}}]\tilde{\mathbf{v}}, \tag{2.42}$$

która określa związek pomiędzy punktami w trójwymiarowej rzeczywistości, a punktami obserwowanymi na obrazach z kamer.

Załóżmy, że dysponujemy zbiorem punktów $\{\tilde{\mathbf{v}}_n\}_{n=1}^N$ z szablonu kalibracyjnego oraz zbiorami odpowiadających im punktów na obrazach z kamer $\{\tilde{\mathbf{v}}_n^{\mathrm{I}}\}_{n=1}^N$. Wtedy dla obrazu $\mathrm{I} \in \mathcal{I}$ z ustalonej kamery możemy oszacować położenie punktu $\tilde{\mathbf{v}}_n^{\mathrm{I}}$ korzystając z równania (2.42) i wyrażając punkt na obrazie jako funkcja $\tilde{\mathbf{v}}^{\mathrm{I}}(\mathbf{A}, \mathbf{R}_{\mathrm{I}}, \mathbf{u}_{\mathrm{I}}, \tilde{\mathbf{v}}_n)$. W warunkach idealnych pomiarów tak uzyskany punkt powinien pokryć się z punktem $\tilde{\mathbf{v}}_n^{\mathrm{I}}$. W praktyce to się nie zdarza, ponieważ pomiary obciążone są losowym szumem. W związku z tym staramy się jak najlepiej dopasować obserwacje do wartości uzyskanych z (2.42) poprzez minimalizowanie następującego kryterium:

$$\sum_{\mathrm{I}\in\mathcal{I}} \sum_{n=1}^{N} \|\tilde{\mathbf{v}}_n^{\mathrm{I}} - \tilde{\mathbf{v}}^{\mathrm{I}}(\mathbf{A}, \mathbf{R}_{\mathrm{I}}, \mathbf{u}_{\mathrm{I}}, \tilde{\mathbf{v}}_n)\|^2. \tag{2.43}$$

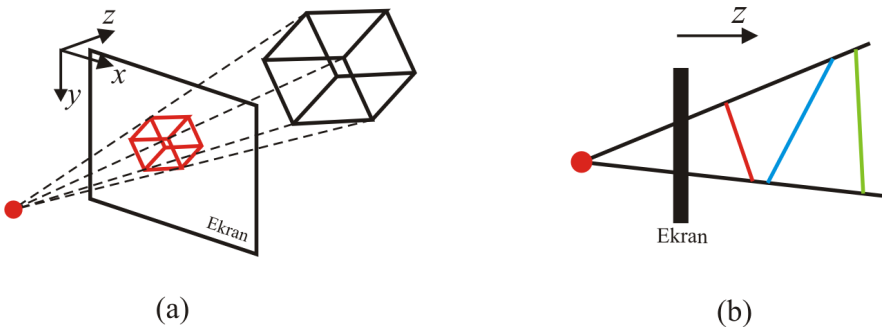
Zmiennymi decyzyjnymi, ze względu na które przeprowadzany jest proces optymalizacji, są niezerowe parametry macierzy $\mathbf{A}$, składowe wektorów przesunięć $\mathbf{u}_{\mathrm{I}}$ oraz parametry, budujące macierze rotacji $\mathbf{R}_{\mathrm{I}}$ zgodnie z formułą Rodriguesa [136] (po trzy parametry dla każdej z kamer). Problem optymalizacji (2.43) jest przykładem nieliniowego zadania najmniejszych kwadratów i może być rozwiązany z użyciem algorytmu Levenberga-Marquardta [115], który jest dedykowany do problemów tego typu. W celu uzyskanie wysokiej jakości rozwiązania istotny jest dobór początkowych wartości zmiennych decyzyjnych. W [175] została opisana procedura pozwalająca ustalić początkowe wartości parametrów blisko rozwiązania optymalnego.

Na koniec warto zauważyć, że w problemie (2.43) zakładamy, że macierz parametrów wewnętrznych $\mathbf{A}$ jest taka sama dla wszystkich kamer. Wprowadzając różne macierze $\mathbf{A}_{\mathrm{I}}$, może to być wprost uogólnione na przypadek, gdzie dysponujemy różnymi typami kamer i każda z nich ma indywidualną macierz parametrów wewnętrznych. Dodatkowo kamera może zniekształcać obraz w sposób nieliniowy. Przykładem takiego zniekształcenia jest zjawisko dystorsji radialnej (ang. radial distortion), charakterystyczne dla słabej klasy obiek-

tywów. Wtedy zależność (2.42) może być rozwinięta o parametry takiego zniekształcenia, które również mogą być oszacowane w procesie minimalizacji funkcji (2.43). Rozszerzenie problemu kalibracji o zjawisko dystorsji jest zaprezentowane w pracy [174].

2.2.2 Projekcja perspektywiczna

Parametry $\mathbf{A}$, $\mathbf{R}_{\mathrm{I}}$, $\mathbf{u}_{\mathrm{I}}$ uzyskane w procesie *kalibracji kamer* mogą być wykorzystane do rzutowania punktu ze sceny na obraz I. Zauważmy, że stosując zależność (2.42) do rozszerzonej reprezentacji punktu $\tilde{\mathbf{v}}$ uzyskamy punkt $\tilde{\mathbf{v}}^{\mathrm{I}}$ jedynie z dokładnością do parametru s. Oznacza to, że leży on na półprostej zaczepionej w punkcie obserwacji i przechodzącej przez ekran (obraz) w punkcie $(\tilde{v}_x^{\mathrm{I}}, \tilde{v}_y^{\mathrm{I}})$ (rysunek 2.4a). Jest to znany fakt z geometrii projekcyjnej (ang. projective geometry), że pojedynczy punkt na obrazie jest reprezentantem całej klasy abstrakcji, do której wchodzą wszystkie punkty z półprostej przechodzącej przez ten punkt [67].



(a) (b)

Rysunek 2.4: Projekcja perspektywiczna. (a) Projekcja obiektu na ekran. (b) Identyczny obraz dla różnych obiektów.

Oczywiście, aby z uzyskanej postaci $s\tilde{\mathbf{v}}^{\mathrm{I}} = (s\tilde{v}_x^{\mathrm{I}}, s\tilde{v}_y^{\mathrm{I}}, s)^{\mathrm{T}}$ otrzymać współrzędne punktu widocznego na obrazie, należy dwie pierwsze składowe podzielić przez trzecią. Operacja ta nazywa się projekcją perspektywiczną (rysunek 2.4a). Formalnie, dla dowolnego punktu $\mathbf{v}$ projekcja perspektywiczna jest postaci:

$$\mathcal{P}(\mathbf{v}) = \left(\frac{v_x}{v_z}, \frac{v_y}{v_z}\right). \tag{2.44}$$

W celu uproszczenia notacji w dalszej części pracy zdefiniujmy pojęcie projekcji punktu $\overline{\mathrm{v}}$ określonego w globalnym układzie współrzędnych na obraz I:

$$\mathcal{P}_{\mathrm{I}}(\overline{\mathrm{v}}) = \mathcal{P}(\mathbf{A}[\mathbf{R}_{\mathrm{I}} \;\; \mathbf{u}_{\mathrm{I}}]\tilde{\mathbf{v}}). \tag{2.45}$$

Korzystając z powyższej definicji, wprowadźmy także pojęcie projekcji elementu sztywnego na obraz:

$$\mathcal{P}_{\mathrm{I}}(\mathcal{V}) = \left\{\mathcal{P}_{\mathrm{I}}(\overline{\mathrm{v}}) : \; \overline{\mathrm{v}} \in \overline{\mathcal{V}}\right\}, \tag{2.46}$$

gdzie $\overline{\mathcal{V}}$ oznacza zbiór punktów należących do elementu sztywnego wyrażonych w globalnych układzie współrzędnych.

Na rysunku 2.4b kolorami czerwonym, zielonym i niebieskim zostały przedstawione trzy różne obiekty, które dają ten sam obraz po zastosowaniu projekcji perspektywicznej. Pokazuje to, w jaki sposób tracona jest informacja o trójwymiarowej rzeczywistości przy obserwowaniu jej na dwuwymiarowych obrazach. W konsekwencji stanowi również o tym, że dla zbioru obrazów $\mathcal{I}$ nie jest możliwe jednoznaczne stwierdzenie, w jakiej konfiguracji znajduje się człowiek.

Rozdział 3

Sformułowanie problemu

W tym rozdziale został formalnie przedstawiony problem *śledzenia ruchu człowieka* (ang. human motion tracking), którego próbę rozwiązania podjęto w pracy. W tym celu został najpierw zaprezentowany problem *estymacji pozy* (ang. pose estimation), który stanowi składową część problemu śledzenia ruchu. Następnie zostały zaproponowane dwie koncepcje mechanizmu śledzenia, jedna znana z literatury oparta na *ukrytym modelu Markowa*, a druga autorska oparta na własnym modelu.

3.1 Problem estymacji pozy

Problem *estymacji pozy* polega na oszacowaniu *wektora stanu* $\mathbf{x}$ na podstawie dostępnego w danej chwili zbioru obrazów $\mathcal{I}$ z wielu zsynchronizowanych kamer. Warto zwrócić uwagę, że żadne dodatkowe pomiary nie są wykonywane. Ponadto jest to problem statyczny, tj. nie uwzględnia się konfiguracji historycznych w celu poprawienia jakości bieżącej predykcji.

W poprzednim rozdziale zostało zauważone, że problem *estymacji pozy* jest źle uwarunkowany (ang. ill-posed). Wynika to z faktu, że obserwowanie trójwymiarowej rzeczywistości na podstawie dwuwymiarowych obrazów powoduje stratę informacji, a w konsekwencji ta sama obserwacja $\mathcal{I}$ może być uzyskana dla wielu różnych konfiguracji $\mathbf{x}$. Zależność w drugą stronę także nie jest jednoznaczna, tj. dostępne obrazy zawierają wiele nadmiarowej informacji, która w rozważanym problemie może być traktowana jako szum. Są to na

przykład dodatkowe obiekty na obrazie wchodzące w skład tła, zmienne oświetlenie, różny ubiór i wygląd człowieka, szum kamery itp. Oznacza to, że ogromnej liczbie różnorodnych obserwacji $\mathcal{I}$ będzie odpowiadał ten sam *wektor stanu*. W konsekwencji występuje tutaj relacja wiele-do-wielu, tj. ani przejście od zbioru obrazów do konfiguracji człowieka, ani odwrotnie, od konfiguracji do zbioru obrazów nie jest jednoznaczne.

Najogólniej zależność pomiędzy *wektorem stanu* i pomiarami możemy opisać łącznym rozkładem prawdopodobieństwa $p(\mathbf{x}, \mathcal{I})$. Rozkład ten jest dla nas nieznany, a jego postać jest dalece nietrywialna. Niemniej przy założeniu znajomości rozkładu możemy rozważać wyznaczenie zależności funkcyjnej $\hat{\mathbf{x}}(\mathcal{I})$, która na podstawie dostępnych obserwacji jednoznacznie zwraca *wektor stanu*. W statystycznej teorii decyzji sprowadza się to do znalezienia funkcji minimalizującej ustalony funkcjonał ryzyka (ang. risk) [15, 17, 35, 152, 165]:

$$R[\hat{\mathbf{x}}] = \iint L(\mathbf{x}, \hat{\mathbf{x}}) p(\mathbf{x}, \mathcal{I}) \mathrm{d}\mathbf{x} \mathrm{d}\mathcal{I}. \tag{3.1}$$

W powyższej definicji $L(\mathbf{x}, \hat{\mathbf{x}})$ oznacza funkcję straty (ang. loss function). W pracy rozważa się dwie następujące funkcje straty, które jednocześnie są najczęściej spotykane w praktyce:

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2}\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \quad \text{(kwadratowa funkcja straty)}, \tag{3.2}$$

$$L(\mathbf{x}, \hat{\mathbf{x}}) = -\delta(\mathbf{x} - \hat{\mathbf{x}}) \quad \text{(minus delta Diraca)}. \tag{3.3}$$

Pierwsza funkcja wymusza, aby podjęta decyzja minimalizowała błąd średniokwadratowy, a co za tym idzie gwarantowała, że podejmiemy optymalną decyzję w sensie średnim. Wstawiając (3.2) do funkcjonału (3.1), a następnie minimalizując go z użyciem metod rachunku wariacyjnego ze względu na funkcję $\hat{\mathbf{x}}(\mathcal{I})$ otrzymujemy:

$$\hat{\mathbf{x}}(\mathcal{I}) = \mathbb{E}[\mathbf{x}|\mathcal{I}] = \int \mathbf{x} p(\mathbf{x}|\mathcal{I}) \mathrm{d}\mathbf{x}. \tag{3.4}$$

Zauważmy zatem, że do podjęcia optymalnej decyzji wystarcza nam znajomość warunkowego rozkładu $p(\mathbf{x}|\mathcal{I})$, zamiast rozkładu łącznego $p(\mathbf{x}, \mathcal{I})$. Warto zaznaczyć, że rozkład warunkowy $p(\mathbf{x}|\mathcal{I})$ jest zazwyczaj wielomodalny, tj. posiada wiele maksimów lokalnych. Wynika to z faktu, że dla danego pomiaru $\mathcal{I}$ istnieje wiele możliwych *wektorów stanu*. W konsekwencji decyzja (3.4) polegająca na uśrednieniu po wszystkich możliwych konfiguracjach może zwracać *wektor stanu*, którego prawdopodobieństwo wystąpienia jest niskie, mimo iż minimalizuje on średniokwadratowy błąd.

Druga funkcja straty w postaci (3.3) powoduje, że wyróżniona zostanie tylko jedna konfiguracja x pokrywającej się z decyzją. Korzystając z elementarnych własności delty Diraca można pokazać, że decyzja minimalizująca ryzyko (3.1) z funkcją straty (3.3) ma następującą postać:

$$\hat{\mathbf{x}}(\mathcal{I}) = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathcal{I}). \tag{3.5}$$

W ten sposób wybieramy *wektor stanu* $\hat{\mathbf{x}}$, który maksymalizuje warunkowy rozkład $p(\mathbf{x}|\mathcal{I})$. Inaczej mówiąc, wybieramy estymator maksymalnego a posteriori (ang. MAP estimation). Ze względu na fakt, że warunkowy rozkład jest wielomodalny, istnieje niebezpieczeństwo, że *wektor stanu* uzyskany poprzez zastosowanie (3.5) będzie odległy od rzeczywistego *wektora stanu*, pomimo faktu, że jego prawdopodobieństwo jest najwyższe.

Formalnie mówiąc, problem *estymacji pozy* polega na wyznaczeniu estymatora $\hat{\mathbf{x}}$ zgodnie z regułą (3.4) lub (3.5). Zauważmy, że w obu przypadkach wymagana jest znajomość warunkowego rozkładu $p(\mathbf{x}|\mathcal{I})$. W rzeczywistości rozkład ten ma bardzo skomplikowaną postać i próba zamodelowania go jest kluczowym elementem do rozwiązania problemu.

Zgodnie z podziałem zaproponowanym w rozdziale 1.2, koncepcje modelowania rozkładu $p(\mathbf{x}|\mathcal{I})$ można podzielić na dwie grupy:

1. *Modele dyskryminacyjne* (ang. discriminative models). Podejście to polega na bezpośrednim zaproponowaniu postaci modelu warunkowego.

2. *Modele generujące* (ang. generative models). W tym podejściu wykorzystuje się *twierdzenie Bayesa* do odwrócenia warunkowania w rozkładzie prawdopodobieństwa:

   $$p(\mathbf{x}|\mathcal{I}) \propto p(\mathcal{I}|\mathbf{x})p(\mathbf{x}). \tag{3.6}$$

   Następnie modeluje się niezależnie rozkład $p(\mathcal{I}|\mathbf{x})$, zwany *modelem wiarygodności* (ang. image likelihood) oraz rozkład a priori na *wektory stanu* $p(\mathbf{x})$. Warto zauważyć, że nie ma konieczności modelowania rozkładu brzegowego $p(\mathcal{I})$, gdyż jest to jedynie czynnik normujący i może być on wyznaczony z następującej zależności:

   $$p(\mathcal{I}) = \int p(\mathcal{I}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{3.7}$$

Rozważane w pracy podejście zalicza się grupy *modelowania generującego*.

3.2 Problem śledzenia ruchu człowieka

Rozważmy teraz problem odtworzenia całej trajektorii, czyli sekwencji *wektorów stanu* $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, bazując na sekwencji pomiarów $\mathcal{I}_{1:T} = \{\mathcal{I}_1, \ldots, \mathcal{I}_T\}$, która odpowiada obrazom pobranym ze wszystkich kamer w kolejnych momentach czasu $1, \ldots, T$.

Podobnie, jak w przypadku problemu *estymacji pozy*, istnieje nieznany dla nas rozkład prawdopodobieństwa $p(\mathbf{x}_{1:T}, \mathcal{I}_{1:T})$ oraz optymalna decyzja dotycząca trajektorii $\hat{\mathbf{x}}_{1:T}$ może być uzyskana poprzez zastosowanie formuł analogicznych do (3.4) i (3.5) dla rozkładu warunkowego $p(\mathbf{x}_{1:T}|\mathcal{I}_{1:T})$. Oczywiście postać tego rozkładu będzie dużo bardziej skomplikowana od rozkładu dla pojedynczego *wektora stanu* $p(\mathbf{x}|\mathcal{I})$.

Stosując podejście generujące, staramy się modelować $p(\mathbf{x}_{1:T}, \mathcal{I}_{1:T}) = p(\mathcal{I}_{1:T}|\mathbf{x}_{1:T})p(\mathbf{x}_{1:T})$ poprzez rozbicie rozkładu łącznego na wiarygodność obrazów i rozkład a priori dla trajektorii. Wprowadźmy następujące założenia dotyczące postaci rozkładów:

1. Rozkład *wektora stanu* w chwili t zależy jedynie od *wektora stanu* w chwili $t-1$. To założenie implikuje następującą warunkową niezależność:

$$\mathbf{x}_t \perp \mathbf{x}_{1:t-2} \mid \mathbf{x}_{t-1}. \tag{3.8}$$

   Zauważmy, że dla dowolnej chwili t *wektor stanu* jest warunkowo niezależny od wszystkich pozostałych *wektorów stanu* za wyjątkiem $\mathbf{x}_{t-1}$ do tej chwili. Prowadzi to do następującej faktoryzacji rozkładu a priori na trajektorię:

$$\begin{aligned} p(\mathbf{x}_{1:T}) &= p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) \\ &= p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}). \end{aligned} \tag{3.9}$$

2. Bieżąca obserwacja $\mathcal{I}_t$ zależy jedynie od bieżącego *wektora stanu* $\mathbf{x}_t$. Konsekwencją tego założenia są następujące warunkowe niezależności:

$$\mathcal{I}_t \perp \mathcal{I}_{1:T\setminus\{t\}} \mid \mathbf{x}_t, \tag{3.10}$$

$$\mathcal{I}_t \perp \mathbf{x}_{1:T\setminus\{t\}} \mid \mathbf{x}_t, \tag{3.11}$$

gdzie notacja $\mathcal{I}_{1:T\setminus\{t\}}$ oznacza wszystkie obserwacje za wyjątkiem obserwacji $\mathcal{I}_t$. Powyższe własności prowadzą do następującej faktoryzacji wiarygodności obrazów:

$$p(\mathcal{I}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(\mathcal{I}_t|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(\mathcal{I}_t|\mathbf{x}_t), \tag{3.12}$$

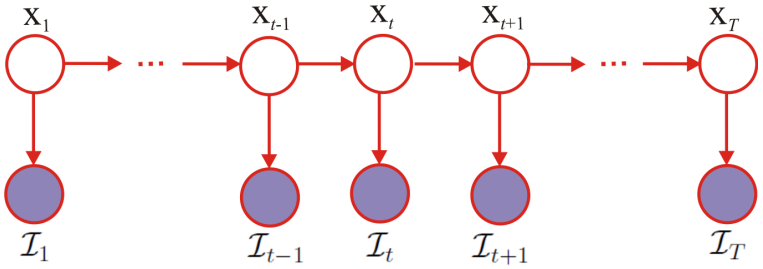
gdzie pierwsza równość jest konsekwencją niezależności obrazów przy znajomości *wektora stanu* (3.10), a druga równość wynika z niezależności pomiaru $\mathcal{I}_t$ od wszystkich *wektorów stanu* za wyjątkiem $\mathbf{x}_t$ (3.11).

Należy podkreślić, że stosowanie powyższych założeń ma na celu uproszczenie problemu i jest jednym z możliwych podejść. Przykładowo w pracy [153] złagodzony został pierwszy warunek i zakłada się, że bieżący wektor stanu zależy od kilku poprzednich.

Składając razem faktoryzacje (3.9) i (3.12) otrzymujemy następującą postać rozkładu łącznego:

$$p(\mathbf{x}_{1:T}, \mathcal{I}_{1:T}) = p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} p(\mathcal{I}_t|\mathbf{x}_t). \tag{3.13}$$

Warto zauważyć, że modele charakteryzujące się powyższą dekompozycją tworzą klasę *ukrytych modeli Markowa* [1] (ang. hidden Markov models). Na rysunku 3.1 została przed-



Rysunek 3.1: Probabilistyczny model grafowy dla ukrytego modelu Markowa.

stawiona reprezentacja klasy ukrytych modeli Markowa przy użyciu *probabilistycznego modelu grafowego* (ang. probabilistic graphical model). Modele grafowe są powszechnie stosowanym narzędziem w *uczeniu maszynowym* do modelowania skomplikowanych rozkładów

[1] Pojęcie klasy ukrytych modeli Markowa nie powinno być mylone z powszechnie stosowanym ukrytym modelem Markowa, gdzie zakłada się, że $\mathbf{x}_t$ przyjmuje wartości ze skończonego zbioru ukrytych stanów.

prawdopodobieństwa [17, 88, 168] i pozwalają przy pomocy prostych graficznych notacji wyrażać zależności pomiędzy zmiennymi losowymi.

Rozważmy łączny rozkład $p(\mathbf{x}_{1:t}, \mathcal{I}_{1:t})$. Stosując te same założenia, jak do postaci rozkładu (3.13) otrzymujemy następującą zależność:

$$p(\mathbf{x}_{1:t}, \mathcal{I}_{1:t}) = p(\mathbf{x}_{1:t-1}, \mathcal{I}_{1:t-1}) p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{3.14}$$

Powyższa postać może być natychmiast uzyskana również z reprezentacji grafowej (rysunek 3.1). Wyznaczmy teraz rozkład warunkowy na sekwencję *wektorów stanu* do momentu t:

$$\begin{aligned} p(\mathbf{x}_{1:t}|\mathcal{I}_{1:t}) &= \frac{p(\mathbf{x}_{1:t-1}, \mathcal{I}_{1:t-1}) p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathcal{I}_{1:t})} \\ &= \frac{p(\mathbf{x}_{1:t-1}, \mathcal{I}_{1:t-1})}{p(\mathcal{I}_{1:t-1})} \frac{p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \\ &= p(\mathbf{x}_{1:t-1}|\mathcal{I}_{1:t-1}) \frac{p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})}. \end{aligned} \tag{3.15}$$

Korzystając z postaci rozkładu warunkowego możemy wyznaczyć rozkład a posteriori na pojedynczy *wektor stanu* w momencie t, bazując na wszystkich dotychczasowych obserwacjach $\mathcal{I}_{1:t}$, poprzez wycałkowanie pozostałych *wektorów stanu*:

$$\begin{aligned} p(\mathbf{x}_t|\mathcal{I}_{1:t}) &= \int p(\mathbf{x}_{1:t}|\mathcal{I}_{1:t}) \mathrm{d}\mathbf{x}_{1:t-1} \\ &= \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{1:t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{1:t-1} \\ &= \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) \int p(\mathbf{x}_{1:t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{1:t-2} \mathrm{d}\mathbf{x}_{t-1} \\ &= \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \\ &\propto p(\mathcal{I}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1}. \end{aligned} \tag{3.16}$$

W ostatnim przekształceniu czynnik normujący $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$ został pominięty, ponieważ podobnie jak w przypadku (3.6) może on być wyznaczony korzystając z pozostałych rozkładów:

$$p(\mathcal{I}_t|\mathcal{I}_{1:t-1}) = \int p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1:t}. \tag{3.17}$$

Zauważmy, że równanie (3.16) wyznacza rekurencyjną zależność pomiędzy rozkładami $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ i $p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$. Innymi słowy, jest to procedura, która pozwala uaktualnić wiedzę o *wektorze stanu* $\mathbf{x}_t$, bazując na dotychczasowej wiedzy oraz na bieżącej obserwacji $\mathcal{I}_t$. Procedura ta nazywa się *filtrowaniem* (ang. filtering) [46]. Powszechnie znanym algorytmem filtrującym jest *filtr Kalmana* (ang. Kalman filter) [17, 82], który zakłada, że wszystkie postaci rozkładów w (3.16) są rozkładami normalnymi. Pomimo licznych zastosowań tej metody, nie sprawdza się ona w problemie śledzenia ruchu człowieka, ze względu na jednomodalny charakter rozkładów normalnych i liniową dynamikę.

Formalnie, problemem *śledzenia ruchu człowieka* sprowadza się do oszacowania trajektorii *wektorów stanu* $\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_T\}$, gdzie pojedynczy $\hat{\mathbf{x}}_t$ jest wyznaczany w procedurze podejmowania decyzji na podstawie warunkowego rozkładu $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ poprzez zastosowanie jednej z reguł decyzyjnych (3.4) lub (3.5). Warunkowy rozkład prawdopodobieństwa $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ jest wyznaczany w procesie *filtrowania* (3.16).

Warto zauważyć, że w przeciwieństwie do problemu *estymacji pozy*, *śledzenie ruchu człowieka* jest problemem dynamicznym, tj. w procedurze *filtrowania* uwzględniona jest wiedza o poprzednim *wektorze stanu*. Konsekwencją tej rekurencji jest to, że rozkład *wektora stanu* $\mathbf{x}_t$ zależy od wszystkich dotychczasowych obserwacji, a nie jak w przypadku *estymacji pozy*, jedynie od bieżącej.

W celu rozwiązania sformułowanego problemu *śledzenia ruchu człowieka*, należy rozstrzygnąć następujące kwestie:

1. Zaproponować model rozkładu a priori na początkowy *wektor stanu* $p(\mathbf{x}_1)$. W pracy przyjmuje się założenie, że początkowy stan jest znany i wynosi $\hat{\mathbf{x}}_1$. Wtedy formalnie rozkład a priori jest postaci:
$$p(\mathbf{x}_1) = \delta(\mathbf{x}_1 - \hat{\mathbf{x}}_1). \tag{3.18}$$
Niemniej stworzenie procedury wyznaczania tego rozkładu jest ciekawym zagadnieniem i powinno być rozumiane jako problem automatycznej inicjalizacji systemu śledzącego.

2. Stworzyć *model dynamiki* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, który będzie determinował, w które fragmenty przestrzeni stanów może przesunąć się śledzona postać, jeśli znajdowała się w stanie $\mathbf{x}_{t-1}$. *Modele dynamiki* zostały zaproponowane w rozdziale 6.

3. Zdefiniować *model wiarygodności* obrazu $p(\mathcal{I}_t|\mathbf{x}_t)$, który będzie pozwalał ocenić na ile prawdopodobne jest, że mogliśmy zaobserwować obraz $\mathcal{I}_t$, jeśli założymy, że stan $\mathbf{x}_t$ jest prawdziwy. *Modele wiarygodności* obrazu zostały zaproponowane w rozdziale 5.

4. Zaproponować metodę, która będzie pozwalała na wyznaczenie całki

$$p(\mathbf{x}_t|\mathcal{I}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}, \tag{3.19}$$

   która występuje w procedurze *filtrowania* (3.16). W literaturze związanej z metodami *filtrowania*, nazywa się to krokiem predykcji (ang. prediction step). Odpowiednia procedura bazująca na *filtrze cząsteczkowym* (ang. particle filter) została zaproponowana w rozdziale 4.1.

5. Określić metodę, która pozwoli poprawić rozkład wektora stanu o nową obserwację:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathcal{I}_{1:t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})}. \tag{3.20}$$

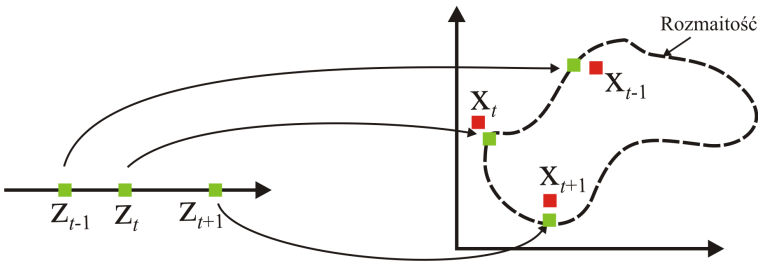
   Procedura ta nazywa się krokiem aktualizacji (ang. update step), a jej trudność polega na wyznaczeniu czynnika normującego (3.17). Ponownie rozwiązane to zostało z użyciem *filtra cząsteczkowego*.

3.2.1 Dynamika w pobliżu niskowymiarowej rozmaitości

Badanie prowadzone nad *śledzeniem ruchu człowieka* pokazały, że zasadniczą część przestrzeni stanów stanowi obszar niedopuszczalnych konfiguracji ludzkiego ciała. Dodatkowo podczas charakterystycznych ruchów (chodzenie, bieganie) wszystkie składowe *wektora stanu* wykazują się silną korelacją, która zmienia się w zależności od stanu, w którym znajduje się człowiek. Prowadzi to do spostrzeżenia, że rzeczywiste trajektorie ruchu rozkładają się w pobliżu niskowymiarowych *rozmaitości* (ang. manifold) [19, 33, 48, 63, 66, 71, 77, 91, 94, 97, 108, 155, 161, 163, 164, 169].

W konsekwencji skomplikowane staje się zamodelowanie rozkładu dynamiki $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, w taki sposób, aby uwzględniał on nieznaną strukturę rozmaitości, w pobliżu której odbywa się ruch.

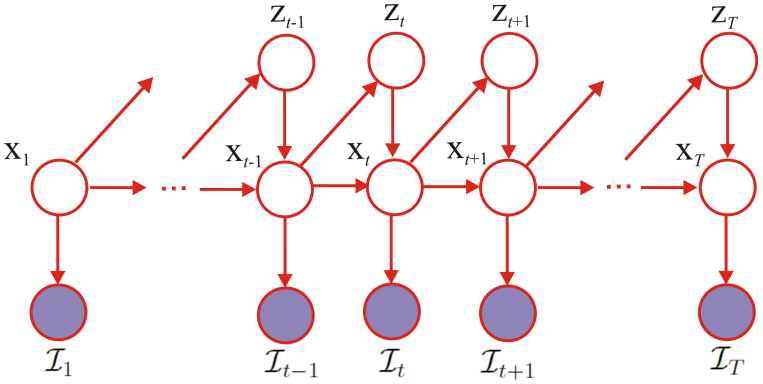
Załóżmy, że z *rozmaitością* związany jest pewien układ współrzędnych i dowolny punkt na niej opisuje zmienna $\mathbf{z}$ (rysunek 3.2). Należy zwrócić uwagę, że wymiar zmiennej $\mathbf{z}$ jest



Rysunek 3.2: Układ współrzędnych na niskowymiarowej rozmaitości.

istotnie mniejszy od wymiaru *wektora stanu* $\mathbf{x}$. Wtedy możemy przyjąć założenie, że dynamika odbywa w pobliżu *rozmaitości*, która zadaje pewien „schemat" w sposobie poruszania się człowieka, a konkretna trajektoria jedynie lokalnie odchyla się od niej. Na rysunku 3.2 realizacje *wektorów stanu* pochodzących z ustalonej trajektorii zostały zaznaczone na czerwono.

Korzystając z *probabilistycznych modeli grafowych* [17, 88, 168], możemy w łatwy sposób uwzględnić zmienne $\mathbf{z}_t$ w opisie prawdopodobieństwa. Na rysunku 3.3 został przed-



Rysunek 3.3: Probabilistyczny model grafowy uwzględniający strukturę niskowymiarowej rozmaitości.

stawiony *probabilistyczny model grafowy*, który pokazuje, w jaki sposób dekomponuje się rozkład łączny dla całej trajektorii $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T})$. Warto zauważyć, że bieżący *wektor stanu* $\mathbf{x}_t$ determinuje przyszły *wektor stanu* i przyszłą współrzędną na *rozmaitości* $\mathbf{z}_{t+1}$, która dodatkowo wpływa na następny *wektor stanu* $\mathbf{x}_{t+1}$. Powoduje to, że do wyznaczenia

przyszłego stanu uwzględniona jest informacja o stanie poprzednim i o położeniu na *rozmaitości*. Zbliżony model został użyty w pracy [153], z tą różnicą, że zależność pomiędzy stanem i współrzędną na *rozmaitości* opisana jest łącznym rozkładem $p(\mathbf{x}_t, \mathbf{z}_t)$. Przykładem innego podejścia jest praca [169], gdzie dynamika zadana jest bezpośrednio na *rozmaitości*, tj. występują rozkłady $p(\mathbf{z}_t|\mathbf{z}_{t-1})$.

Korzystając z modelu grafowego możemy wyrazić łączny rozkład prawdopodobieństwa:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T}) = p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{x}_{t-1}) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) \prod_{t=1}^{T} p(\mathcal{I}_t|\mathbf{x}_t). \tag{3.21}$$

Analogicznie, jak w przypadku wzoru (3.14) możemy wyznaczyć łączny rozkład do momentu t:

$$p(\mathbf{x}_{1:t}, \mathbf{z}_{1:t}, \mathcal{I}_{1:t}) = p(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathcal{I}_{1:t-1}) p(\mathbf{z}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathcal{I}_t|\mathbf{x}_t). \tag{3.22}$$

Następnie wyznaczamy rozkład warunkowy, bazując na wszystkich obserwacjach do bieżącego momentu, podobnie jak w zależności (3.15):

$$p(\mathbf{x}_{1:t}, \mathbf{z}_{1:t}|\mathcal{I}_{1:t}) = p(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}|\mathcal{I}_{1:t-1}) \frac{p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{z}_t|\mathbf{x}_{t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \tag{3.23}$$

Ostatecznie wycałkowując wszystkie zmienne poza bieżącym stanem, otrzymujemy rozkład a posteriori na $\mathbf{x}_t$, analogicznie jak w przypadku (3.16):

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \iint p(\mathbf{z}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \mathrm{d}\mathbf{z}_t, \tag{3.24}$$

gdzie czynnik normujący może być wyznaczony z zależności:

$$p(\mathcal{I}_t|\mathcal{I}_{1:t-1}) = \iint p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{z}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1:t} \mathbf{z}_t. \tag{3.25}$$

Otrzymaliśmy w ten sposób procedurę *filtrowania*, która uwzględnia istnienie niskowymiarowej *rozmaitości*, w pobliżu której odbywa się ruch. Warto zwrócić uwagę, że jeśli możliwe byłoby wycałkowanie po zmiennej $\mathbf{z}_t$, to wtedy otrzymujemy *model dynamiki* na *wektorze stanu*, bez uwzględniania położenia na *rozmaitości* wprost:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{z}_t|\mathbf{x}_{t-1}) \mathrm{d}\mathbf{z}_t. \tag{3.26}$$

Powoduje to zredukowanie problemu *filtrowania* (3.24) do procedury (3.16). Zazwyczaj jednak powyższa całka nie ma analitycznego rozwiązania.

Reasumując, aby rozwiązać problem *śledzenia ruchu* z uwzględnieniem niskowymiarowej *rozmaitości* należy:

1. Zaproponować modele $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ oraz $p(\mathbf{z}_t|\mathbf{x}_{t-1})$, zamiast *modelu dynamiki* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Zostało to przedstawione w rozdziale 6.2.

2. Określić metodę, która pozwoli wykonać procedurę *filtrowania* zdefiniowaną zależnością (3.24). W rozdziale 4.3 została zaproponowana modyfikacja *filtra cząsteczkowego*, rozwiązującego kroki predykcji i aktualizacji dla problemu (3.24).

3.3 Ocena jakości rozwiązania

Jakość zaproponowanych modeli, tj. *modeli dynamiki* i *wiarygodności*, a także skuteczność metod *filtrowania* (3.16) i (3.24) jest oceniana poprzez jakość uzyskanego *wektora stanu* $\hat{\mathbf{x}}_t$ na wyjściu systemu. Bazując na rzeczywistych *wektorach stanu* $\mathbf{x}_t$, które mogą być otrzymane przy użyciu *systemu MOCAP*, możemy porównać na ile otrzymane rozwiązanie jest poprawne.

Ze względu na fakt, że w literaturze definicja *wektora stanu* różni się pod względem przyjętej liczby stopni swobody, jak również stosowanej reprezentacji obrotów, dlatego istotne jest, aby zaproponowana metoda oceny jakości była niezależna od tych różnic. Poniżej została opisana technika zaproponowana w [10, 140], która jest powszechnie stosowana do oceny jakości algorytmów śledzących [16, 19, 24, 34, 41, 59, 63, 94, 97, 108, 120, 142, 128, 153].

Zakładamy, że człowiek jest reprezentowany przy pomocy *drzewa kinematycznego* opisanego w rozdziale 2.1.2. Wtedy możemy go traktować jako zbiór elementów sztywnych $\{\mathcal{V}_0, \ldots, \mathcal{V}_K\}$. Wyróżnijmy grupę punktów testowych $\mathbf{w}$ takich, że każdy punkt należy do wybranego elementu sztywnego. Intuicyjnie punkty te możemy traktować jako „pseudoznaczniki" przyklejone na stałe do człowieka. Wprowadźmy następujące oznaczenie na zbiór wszystkich wybranych punktów:

$$\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}. \tag{3.27}$$

Przez $\mathbf{w}_i(\mathbf{x})$ oznaczmy położenie i-tego punktu testowego w globalnym układzie współrzędnych uzyskane na podstawie *wektora stanu* $\mathbf{x}$. Może ono być wyznaczone poprzez wyliczenie odpowiednich macierzy rotacji z *wektora stanu*, z użyciem formuły (2.12) lub

(2.19) w zależności od reprezentacji obrotu, a następnie rekurencyjne zastosowanie wzorów (2.33) i (2.31).

Dla danej sekwencji rzeczywistych *wektorów stanu* $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ oraz sekwencji estymat będącej wynikiem działania systemu $\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_T\}$ stosowana miara oceny jakości jest następującej postaci:

$$\mathrm{err}(\hat{\mathbf{x}}_{1:T}) = \frac{1}{TM} \sum_{t=1}^{T} \sum_{i=1}^{M} \|\mathbf{w}_i(\mathbf{x}_t) - \mathbf{w}_i(\hat{\mathbf{x}}_t)\|. \tag{3.28}$$

Oznacza to, że weryfikacja działania algorytmu do *śledzenia ruchu człowieka* polega na zmierzeniu, o ile średnio odległe są „pseudoznaczniki" od ich prawidłowego położenia.

Warto wspomnieć, że dodatkową zaletą takiego sposobu oceny jakości, jest fizyczna interpretacja błędu, jako średniej odległości wyrażonej w konkretnych jednostkach, np. w milimetrach.

Rozdział 4

Metody filtrowania

W poniższym rozdziale zaprezentowane zostały dwie metody filtrowania znane z literatury do rozwiązania problemu (3.16): *filtr cząsteczkowy* i *wyżarzany filtr cząsteczkowy* oraz autorska metoda do rozwiązania problemu (3.24) – *filtr cząsteczkowy* uwzględniający niskowymiarową rozmaitość.

4.1 Filtr cząsteczkowy

Filtry cząsteczkowe (ang. particle filter) stanowią grupę algorytmów rozwiązujących zadanie *filtrowania*, bazującą na statystycznych *metodach Monte Carlo* [17, 113, 129]. Są one szczególnym przypadkiem szerszej klasy algorytmów, znanych pod nazwą *sekwencyjne Monte Carlo* (ang. Sequential Monte Carlo).

Ogólna idea stojąca za *filtrami cząsteczkowymi* bazuje na pomyśle, aby aproksymować ciągły rozkład prawdopodobieństwa $p(\mathbf{x}_t|\mathcal{I}_{1:t})$, opisany zależnością (3.16), rozkładem dyskretnym skoncentrowanym w N punktach. Taka forma aproksymacji bardzo dobrze sprawdza się w licznych problemach praktycznych, w tym w problemach śledzenia obiektów na obrazach wideo, gdzie pierwszy raz została zastosowana w pracy [75] pod nazwą *algorytm CONDENSATION*. Odtąd *filtry cząsteczkowe* z różnymi modyfikacjami są powszechnie używane do *śledzenia ruchu człowieka* [24, 26, 32, 41, 44, 66, 86, 94, 97, 119, 120, 128, 131, 139, 140, 153, 155].

W ostatnich latach zostało pokazanych wiele teoretycznych własności, przemawiających

za skutecznością *filtrów cząsteczkowych*, m.in. zbieżność prawie wszędzie aproksymacji do rzeczywistego rozkładu przy $N \to \infty$ [38], zbieżności estymatorów dla pewnych klas nieograniczonych funkcji [73, 74], oszacowania jakości estymatorów z użyciem centralnego twierdzenia granicznego (ang. central limit theorem) [36, 46].

Zanim przedstawiony zostanie algorytm *filtrowania* do *śledzenia ruchu człowieka* oparty na *filtrze cząsteczkowym*, wprowadźmy kilka koncepcji, które uzasadniają jego działanie. Załóżmy, że dysponujemy próbą wygenerowaną z pewnego rozkładu, tj.

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \sim p(\mathbf{x}). \tag{4.1}$$

Wtedy rozkład $p(\mathbf{x})$ może być aproksymowany następującym rozkładem dyskretnym, skoncentrowanym w wygenerowanych punktach:

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\mathbf{x} - \mathbf{x}^{(n)}). \tag{4.2}$$

Łatwo pokazać, że (4.2) jest prawidłowym rozkładem prawdopodobieństwa. Wtedy wartość oczekiwana z dowolnej funkcji może być przybliżona następującym wyrażeniem:

$$\begin{aligned} \mathbb{E}[f(\mathbf{x})] &= \int f(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &\approx \int f(\mathbf{x}) \hat{p}(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{N} \sum_{n=1}^{N} \int f(\mathbf{x}) \delta(\mathbf{x} - \mathbf{x}^{(n)}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}^{(n)}). \end{aligned} \tag{4.3}$$

Można pokazać, że powyższy estymator wartości oczekiwanej jest nieobciążony (ang. unbiased), a jego wariancja jest rzędu $O(1/N)$ i nie zależy od wymiaru wektora $\mathbf{x}$.

Rozważmy teraz problem *estymacji pozy* sformułowany w rozdziale 3.1. Chcemy aproksymować rozkład a posteriori $p(\mathbf{x}|\mathcal{I})$, korzystając z przybliżenia rozkładu a priori w postaci

(4.2). Wtedy mamy:

$$\begin{aligned} p(\mathbf{x}|\mathcal{I}) &= \frac{p(\mathcal{I}|\mathbf{x})p(\mathbf{x})}{p(\mathcal{I})} \\ &\approx \frac{p(\mathcal{I}|\mathbf{x})\hat{p}(\mathbf{x})}{p(\mathcal{I})} \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{p(\mathcal{I}|\mathbf{x})}{p(\mathcal{I})}\delta(\mathbf{x}-\mathbf{x}^{(n)}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{\tilde{\pi}(\mathbf{x}^{(n)})}{p(\mathcal{I})}\delta(\mathbf{x}-\mathbf{x}^{(n)}), \end{aligned} \tag{4.4}$$

gdzie zdefiniowano $\tilde{\pi}(\mathbf{x}^{(n)}) = p(\mathcal{I}|\mathbf{x}^{(n)})$. Korzystając z tej samej aproksymacji dla czynnika normującego $p(\mathcal{I})$ i używając zależność (4.3) otrzymujemy:

$$\begin{aligned} p(\mathcal{I}) &\approx \int p(\mathcal{I}|\mathbf{x})\hat{p}(\mathbf{x}) \\ &= \frac{1}{N}\sum_{n=1}^{N} p(\mathcal{I}|\mathbf{x}^{(n)}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\tilde{\pi}(\mathbf{x}^{(n)}). \end{aligned} \tag{4.5}$$

Podstawiając wówczas (4.5) do (4.4) dostajemy przybliżenie rozkładu a posteriori o postaci:

$$\hat{p}(\mathbf{x}|\mathcal{I}) = \sum_{n=1}^{N}\pi(\mathbf{x}^{(n)})\delta(\mathbf{x}-\mathbf{x}^{(n)}), \tag{4.6}$$

gdzie współczynniki $\pi(\mathbf{x}^{(n)})$ zostały zdefiniowane poprzez unormowanie współczynników $\tilde{\pi}(\mathbf{x}^{(n)})$:

$$\pi(\mathbf{x}^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}^{(n)})}{\sum_{j=1}^{N}\tilde{\pi}(\mathbf{x}^{(j)})}. \tag{4.7}$$

Metoda aproksymacji rozkładu (4.6) jest szczególnym przykładem tzw. *próbkowania znaczącego* (ang. importance sampling). Ponadto pełna informacja o rozkładzie (4.6) zawarta jest w następującym zbiorze:

$$\mathcal{X}^{\pi} = \{(\mathbf{x}^{(n)}, \pi(\mathbf{x}^{(n)}))\}_{n=1}^{N}, \tag{4.8}$$

gdzie pojedyncza para $(\mathbf{x}^{(n)}, \pi(\mathbf{x}^{(n)}))$ nazywana jest cząsteczką (ang. particle).

Warto zauważyć, że postać (4.6) jest bardzo wygodna z praktycznego punktu widzenia, ponieważ wprost możemy wyliczyć oszacowanie *wektora stanu*, stosując jedną z reguł decyzyjnych. Dla reguły (3.4) mamy:

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbb{E}[\mathbf{x}|\mathcal{I}] \\ &\approx \sum_{n=1}^{N} \pi(\mathbf{x}^{(n)})\mathbf{x}^{(n)}.\end{aligned} \tag{4.9}$$

Alternatywnie, dla reguły wyznaczającej estymator maksymalnego a posteriori (3.5) otrzymujemy:

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathcal{I}) \\ &\approx \arg\max_{n} \pi(\mathbf{x}^{(n)}).\end{aligned} \tag{4.10}$$

Rozważmy teraz problem *śledzenia ruchu człowieka*, gdzie równanie (3.16) zostało przekształcone do następującej postaci poprzez założenie, że istnieje analityczne rozwiązanie występującej w nim całki:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathcal{I}_{1:t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})}. \tag{4.11}$$

Zakładamy, że posiadamy próbę z rozkładu $p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$:

$$\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)} \sim p(\mathbf{x}_t|\mathcal{I}_{1:t-1}). \tag{4.12}$$

Wtedy możemy wykonać aproksymację dla rozkładu $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ analogiczną do (4.4), otrzymując:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \frac{1}{N}\sum_{n=1}^{N} \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})}\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{4.13}$$

a następnie stosując takie samo przybliżenie dla czynnika normującego $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$, jak w przypadku (4.5), dostajemy ostateczną postać aproksymacji dla rozkładu a posteriori na *wektor stanu*:

$$\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t}) = \sum_{n=1}^{N} \pi(\mathbf{x}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}). \tag{4.14}$$

Podobnie jak poprzednio, pełna informacja o powyższym rozkładzie zawarta jest w zbiorze cząsteczek:

$$\mathcal{X}_t^{\pi} = \{(\mathbf{x}_t^{(n)}, \pi(\mathbf{x}_t^{(n)}))\}_{n=1}^{N}, \tag{4.15}$$

a estymaty *wektora stanu* $\hat{\mathbf{x}}_t$ mogą być wyznaczone poprzez zastosowanie jednego z kryteriów decyzyjnych (4.9) lub (4.10).

Pozostaje problem wygenerowania próby (4.12). W tym celu rozważmy następującą dekompozycję rozkładu łącznego $p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$:

$$\begin{aligned} p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) &= p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathcal{I}_{1:t-1}) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \\ &= p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}), \end{aligned} \tag{4.16}$$

gdzie skorzystaliśmy z warunkowej niezależności (3.8). Dzięki powyższej faktoryzacji możemy do wygenerowania próby z rozkładu łącznego zastosować procedurę:

1. Wygeneruj realizację z rozkładu $p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$, tj. $\overline{\mathbf{x}}_{t-1}^{(n)} \sim p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$.

2. Korzystając z uzyskanej wartości, wygeneruj realizację z rozkładu warunkowego, tj. $\mathbf{x}_t^{(n)} \sim p(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$.

Powyższy schemat prowadzi do uzyskania realizacji $(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)})$ z rozkładu łącznego (4.16). W szczególności pojedyncza wartość $\mathbf{x}_t^{(n)}$ jest prawidłową realizacją z rozkładu brzegowego $p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$. W ten sposób posiadając próbę z rozkładu a posteriori w chwili $t-1$, tj. $\overline{\mathbf{x}}_{t-1}^{(1)}, \ldots, \overline{\mathbf{x}}_{t-1}^{(N)} \sim p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$, możemy wygenerować próbę z rozkładu a priori w chwili t, tj. $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)} \sim p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$.

Aby skorzystać z powyższej procedury potrzebujemy próby z rozkładu a posteriori w chwili $t-1$. Możemy ją wygenerować korzystając z dyskretnej aproksymacji rozkładu a posteriori zadanej przez zależność (4.14). Dzięki temu otrzymujemy:

$$\overline{\mathbf{x}}_{t-1}^{(1)}, \ldots, \overline{\mathbf{x}}_{t-1}^{(N)} \sim \hat{p}(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}). \tag{4.17}$$

Powyższy schemat postępowania nazywa się *ponownym próbkowaniem* (ang. resampling).

Podsumowując dotychczasowe rozważania możemy zaproponować procedurę do *śledzenia ruchu człowieka*, która opiera się na przybliżonym rozwiązywaniu zadania *filtrowania* (3.16). Procedura ta nosi nazwę *filtra cząsteczkowego* (ang. particle filter) i została opisana algorytmem 1.

Na rysunku 4.1 została zaprezentowana idea działania *filtra cząsteczkowego*, gdzie zaczynając od próby z rozkładu a posteriori w chwili $t-1$, generujemy próbę z rozkładu a

Algorithm 1: Filtr cząsteczkowy do śledzenia ruchu człowieka

Wejście: Stan początkowy $\mathbf{x}_0$, sekwencja pomiarów $\mathcal{I}_{1:T}$

Wyjście: Sekwencja estymat wektora stanu $\hat{\mathbf{x}}_{1:T}$

Powiel początkowy stan $\mathbf{x}_0$ do zbioru $\overline{\mathcal{X}}_0 = \{\overline{\mathbf{x}}_0^{(1)}, \ldots, \overline{\mathbf{x}}_0^{(N)}\}$;

for $t = 1 : T$ **do**

  Wygeneruj próbę $\mathcal{X}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)}\}$ z modelu dynamiki, gdzie $\mathbf{x}_t^{(n)} \sim p(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;

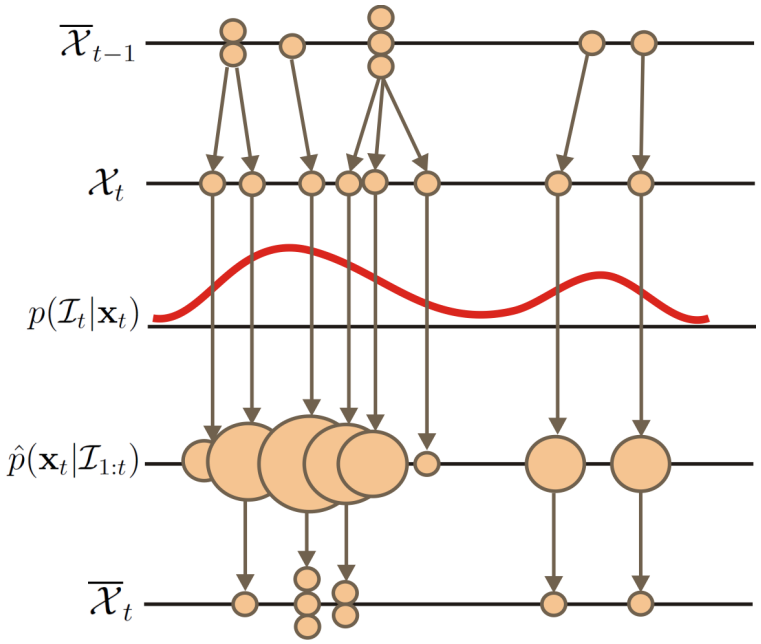
  Wylicz wartości $\tilde{\pi}(\mathbf{x}_t^{(n)})$ korzystając z modelu wiarygodności $p(\mathcal{I}_t|\mathbf{x}_t)$;

  Znormalizuj uzyskane wartości do $\pi(\mathbf{x}_t^{(n)})$ korzystając z zależności (4.7);

  Wyznacz estymatę wektora stanu $\hat{\mathbf{x}}_t$ używając reguły decyzyjnej (4.9) lub (4.10);

  Wygeneruj próbę $\overline{\mathcal{X}}_t = \{\overline{\mathbf{x}}_t^{(1)}, \ldots, \overline{\mathbf{x}}_t^{(N)}\}$ z rozkładu $\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t})$;

end



Rysunek 4.1: Schemat działania filtra cząsteczkowego.

priori w chwili t wykorzystując *model dynamiki* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, następnie wyznaczamy przybliżony rozkład a posteriori w chwili t przy użyciu *modelu wiarygodności* $p(\mathcal{I}_t|\mathbf{x}_t)$ i generujemy z niego próbę. Taka procedura pozwala na sekwencyjne *śledzenie ruchu człowieka* i

uwzględnianie na bieżąco pojawiających się obserwacji.

Na koniec warto przytoczyć kilka istotnych uwag dotyczących *filtra cząsteczkowego*:

1. Istnieje podejście, które nie wymaga *ponownego próbkowania* (ang. resampling) (4.17). Zamiast tego kolejne próby mogą być generowane jedynie z użyciem *modelu dynamiki*, a wagi $\pi(\mathbf{x}_t^{(n)})$ odpowiednio sekwencyjnie poprawiane poprzez uwzględnianie kolejnych obserwacji. Taki schemat postępowania nosi nazwę *sekwencyjnego próbkowania znaczącego* (ang. sequential importance sampling). Można pokazać, że wówczas wariancja estymatora $\hat{\mathbf{x}}_t$ rośnie wykładniczo wraz ze wzrostem t, podczas gdy w przypadku zastosowania procedury *ponownego próbkowania* wariancja rośnie jedynie liniowo [46]. Prowadzi to do tzw. zjawiska degeneracji cząsteczek i szybkiego zaniku prawidłowego śledzenia człowieka, dlatego takie podejście nie jest rozważane w pracy.

2. W opisanym algorytmie zostało przyjęte założenie, że potrafimy wygenerować próbę z *modelu dynamiki* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. W ogólności tak być nie musi i wtedy należy skorzystać z pomocniczego rozkładu, z którego potrafimy generować próbę (np. wielowymiarowego rozkładu normalnego), a następnie wyliczyć dla takiej próby wartość funkcji gęstości $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ w celu ustalenia znaczenia poszczególnych realizacji.

3. *Ponowne próbkowanie* (4.17) może być wykonane poprzez standardową procedurę próbkowania z rozkładu dyskretnego skoncentrowanego w N punktach. Niemniej istnieją metody, które pozwalają na uzyskanie lepszej jakości próby niż przy użyciu standardowej procedury [46].

4. W przypadku *śledzenia ruchu człowieka* pojawia się problem występowania licznych maksimów lokalnych w funkcji gęstości $p(\mathbf{x}_t|\mathcal{I}_{1:t})$. Oznacza to, że zazwyczaj potrzebujemy znaczącą liczbę cząsteczek (kilka tysięcy), aby stosowana aproksymacja prawidłowo oddawała charakter rzeczywistego rozkładu. Wiąże się to z istotnym wzrostem czasu wykonania jednej iteracji, ponieważ pojedyncze wyliczenie funkcji wiarygodności $p(\mathcal{I}_t|\mathbf{x}_t)$ jest zazwyczaj bardzo wymagające obliczeniowo.

5. W literaturze zostały zaproponowane liczne modyfikacje i rozszerzenia *filtrów cząsteczkowych*, m.in. *filtry cząsteczkowe Rao-Blackwella* (ang. Rao-Blackwellized particle

filter), które łączą własności *filtrów Kalmana* i *filtrów cząsteczkowych* [30, 45], *filtry cząsteczkowe* dla nieliniowych dynamik wykorzystujące procesy Gaussa [87], mieszaniny *filtrów cząsteczkowych* [116].

4.2 Wyżarzany filtr cząsteczkowy

Kluczowym problemem w *śledzeniu ruchu człowieka* jest przeszukiwanie wysokowymiarowej przestrzeni stanów, w celu znalezienia prawdopodobnych konfiguracji ciała, które mogą być widoczne na zarejestrowanym zestawie obrazów. W celu uzyskania wysokiej jakości estymat pożądanym zjawiskiem jest, aby możliwie dużo cząsteczek koncentrowało się wokół ekstremów lokalnych funkcji $p(\mathbf{x}_t|\mathcal{I}_{1:t})$. Ze względu na wysoki wymiar przestrzeni stanów, pojawia się tutaj zjawisko *klątwy wymiarowości* (ang. curse of dimensionality), które wnosi, że liczba cząsteczek potrzebnych do pokrycia przestrzeni rośnie wykładniczo wraz ze wzrostem liczby wymiarów. Prowadzi to do problemu koncentrowania się niewielu cząsteczek w pobliżu licznych ekstremów w funkcji gęstości, nawet gdy sumaryczna liczba cząsteczek jest duża. W konsekwencji podstawowa wersja *filtra cząsteczkowego* daje często bardzo niedokładne oszacowania *wektora stanu*.

W nawiązaniu do wspomnianego problemu w pracy [44] został zaproponowany *wyżarzany filtr cząsteczkowy* (ang. annealed particle filter), który wykorzystuje technikę *symulowanego wyżarzania* (ang. simulated annealing) [85] do zwiększenia koncentracji cząsteczek wokół lokalnych maksimów funkcji gęstości. Idea tej metody polega na iteracyjnym „zmniejszaniu temperatury" cząsteczek prowadzącym do ich koncentracji w miejscach o wysokiej wartości funkcji gęstości.

Mówiąc formalnie, dla każdej chwili t, wykonujemy L razy aproksymację rzeczywistego rozkładu a posteriori rozkładem o postaci (4.14), gdzie za każdym razem modyfikujemy go zgodnie z następującą zależnością:

$$\hat{p}_l(\mathbf{x}_t|\mathcal{I}_{1:t}) \propto \hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t})^{\beta_l}, \tag{4.18}$$

gdzie $l = 1, \ldots, L$ i nazywa się warstwą wyżarzania (ang. annealing layer). Parametry $\beta_1, \ldots, \beta_L$ mogą być interpretowane jako odwrotność temperatury, tj. im ich wartość jest niższa, wtedy rozkład ma bardziej jednostajny charakter, a wówczas cząsteczki będą bardziej równomiernie rozłożone po przestrzeni (mają wysoką energię). Natomiast, gdy ich

wartość jest wysoka, wtedy rozkład staje się skoncentrowany wokół punktów modalnych i tam zbierają się cząsteczki. Dodatkowo należy zauważyć, że modyfikacja rozkładu (4.14) zgodnie z zależnością (4.18) prowadzi jedynie do zmiany parametrów wag:

$$\pi_l(\mathbf{x}_{t,l}^{(n)}) \propto \pi(\mathbf{x}_{t,l}^{(n)})^{\beta_l}, \tag{4.19}$$

gdzie normalizacja do postaci $\pi_l(\mathbf{x}_{t,l}^{(n)})$ odbywa się według analogicznej zależności do (4.7). Natomiast $\mathbf{x}_{t,l}^{(n)}$ oznacza realizację wygenerowaną z modelu dynamiki w chwili t i warstwie wyżarzania l.

Do rozstrzygnięcia pozostaje kwestia dobrania parametrów β_l. Najprostszym sposobem jest ich określenie w taki sposób, aby tworzyły ciąg rosnący $\beta_1 < \ldots < \beta_L$, innymi słowy, aby temperatura malała. Okazuje się jednak, że do ich dobrania lepiej jest zastosować następujące kryterium:

$$ESS(\beta_l) = \left(\sum_{n=1}^{N} \pi_l(\mathbf{x}_{t,l}^{(n)}) \right)^{-1}, \tag{4.20}$$

które w literaturze pojawia się pod różnymi nazwami, m.in. efektywna wielkość próby (ang. effective sample size) [46], diagnostyka przetrwania cząsteczek (ang. survival diagnostic) [44, 103]. Wartość tego kryterium może być interpretowana, jako wielkość próby z rzeczywistego rozkładu, która jest potrzebna do uzyskania decyzji o takiej samej jakości, jak decyzja podjęta na podstawie aproksymowanego rozkładu z wagami $\pi_l(\mathbf{x}_{t,l}^{(n)})$. Na tej podstawie możemy uzyskać odsetek efektywnych cząsteczek:

$$\alpha_p(\beta_l) = \frac{ESS(\beta_l)}{N}. \tag{4.21}$$

Wtedy parametry wyżarzania β_l mogą być dobrane tak, aby do następnej warstwy „przechodził" tylko pewien pożądany odsetek α_p cząsteczek, które skupione są w rejonach o wyższym prawdopodobieństwie. Przykładowo, jeśli ustalimy $\alpha_p = 0.5$, wtedy do następnej warstwy efektywnie „przejdzie" połowa cząsteczek. Ostatecznie w celu wyznaczenia parametru β_l należy rozwiązać następujące zadanie optymalizacji:

$$\min_{\beta_l} \quad \Big(\alpha_p - \alpha_p(\beta_l)\Big)^2. \tag{4.22}$$

Ze względu na fakt, że jest to przykład nieliniowego zadania najmniejszych kwadratów, można je efektywnie rozwiązać z użyciem np. algorytmu Levenberga-Marquadta [115].

Dodatkowo, aby cząsteczki miały możliwość skupienia się wokół lokalnych ekstremów funkcji gęstości, należy modyfikować model dynamiki dla każdej warstwy wyżarzania tak, aby wykonywane ruchy były coraz mniejsze. Intuicyjnie można to interpretować, że wraz ze spadkiem temperatury, cząsteczki powinny poruszać się wolniej, gdyż mają mniejszą energię. Zatem kolejne warstwy wyżarzania, będą miały indywidualne modele dynamiki $p_l(\mathbf{x}_t|\mathbf{x}_{t-1})$ takie, że wraz ze wzrostem l wariancja dla tych modeli będzie maleć. W [44] został przedstawiony sposób modyfikacji wariancji, gdzie modelem dynamiki jest dyfuzja Gaussa.

Algorithm 2: Wyżarzany filtr cząsteczkowy do śledzenia ruchu człowieka

Wejście: Stan początkowy $\mathbf{x}_0$, sekwencja pomiarów $\mathcal{I}_{1:T}$

Wyjście: Sekwencja estymat wektora stanu $\hat{\mathbf{x}}_{1:T}$

Powiel początkowy stan $\mathbf{x}_0$ do zbioru $\overline{\mathcal{X}}_{1,0} = \{\overline{\mathbf{x}}_{1,0}^{(1)}, \ldots, \overline{\mathbf{x}}_{1,0}^{(N)}\}$;

for $t = 1 : T$ **do**

  for $l = 1 : L$ **do**

    Wygeneruj próbę $\mathcal{X}_{t,l} = \{\mathbf{x}_{t,l}^{(1)}, \ldots, \mathbf{x}_{t,l}^{(N)}\}$ z modelu dynamiki dla warstwy l, gdzie $\mathbf{x}_{t,l}^{(n)} \sim p_l(\mathbf{x}_t|\overline{\mathbf{x}}_{t,l-1}^{(n)})$;

    Wylicz wartości $\tilde{\pi}(\mathbf{x}_{t,l}^{(n)})$ korzystając z modelu wiarygodności $p(\mathcal{I}_t|\mathbf{x}_t)$;

    Znormalizuj uzyskane wartości do $\pi(\mathbf{x}_{t,l}^{(n)})$ korzystając z zależności (4.7);

    Wyznacz parametr wyżarzania β_l optymalizując kryterium (4.22);

    Wylicz zmodyfikowane wagi $\tilde{\pi}_l(\mathbf{x}_{t,l}^{(n)}) := \pi(\mathbf{x}_{t,l}^{(n)})^{\beta_l}$;

    Znormalizuj zmodyfikowane wagi do $\pi_l(\mathbf{x}_{t,l}^{(n)})$ korzystając z zależności (4.7);

    Wygeneruj próbę $\overline{\mathcal{X}}_{t,l} = \{\overline{\mathbf{x}}_{t,l}^{(1)}, \ldots, \overline{\mathbf{x}}_{t,l}^{(N)}\}$ z rozkładu $\hat{p}_l(\mathbf{x}_t|\mathcal{I}_{1:t})$;

  end

  Wyznacz estymatę wektora stanu $\hat{\mathbf{x}}_t$ używając reguły decyzyjnej (4.9) lub (4.10) na podstawie rozkładu $\hat{p}_L(\mathbf{x}_t|\mathcal{I}_{1:t})$;

  Podstaw $\overline{\mathcal{X}}_{t+1,0} := \overline{\mathcal{X}}_{t,L}$;

end

Podsumowując powyższe rozważania, algorytm 2 przestawia procedurę do *śledzenia ruchu człowieka* opartą na *wyżarzanym filtrze cząsteczkowym*. Warto zauważyć, że zasadniczą modyfikacją w stosunku do podstawowego *filtra cząsteczkowego* (algorytm 1) jest wielo-

krotne generowanie prób $\overline{\mathcal{X}}_{t,l}$ dla pojedynczej chwili t, każdorazowo stosując procedurę wyżarzania. W konsekwencji kolejne zbiory $\overline{\mathcal{X}}_{t,1}, \ldots, \overline{\mathcal{X}}_{t,L}$ są coraz bardziej skoncentrowane wokół ekstremów funkcji $p(\mathbf{x}_t|\mathcal{I}_{1:t})$.

Na koniec warto podkreślić następujące kwestie na temat *wyżarzanego filtra cząsteczkowego*:

1. Częstym efektem, który można zaobserwować podczas stosowania procedury wyżarzania, jest koncentrowanie się cząsteczek wokół tylko jednego dominującego ekstremum. Prowadzi to do niewłaściwej aproksymacji $p(\mathbf{x}_t|\mathcal{I}_{1:t})$, a w konsekwencji do zgubienia się systemu śledzącego. W literaturze odnotowano zarówno takie rezultaty, gdzie *wyżarzany filtr cząsteczkowy* daje lepsze wyniki śledzenia [140], jak i takie, gdzie lepszy jest zwykły *filtr cząsteczkowy* [41].

2. *Wyżarzany filtr cząsteczkowy* potrzebuje zazwyczaj istotnie mniej cząsteczek (kilkaset) niż zwykły *filtr cząsteczkowy*, aby uzyskać porównywalną jakość estymat *wektora stanu*. Niemniej dla każdej chwili t potrzebuje wykonać L razy procedurę, którą *zwykły filtr* wykonuje tylko raz, a w konsekwencji jego złożoność obliczeniowa jest przeważnie większa.

3. Ustalenia wymagają dodatkowe parametry, których nie było w zwykłym *filtrze cząsteczkowym*, jak liczba warstw wyżarzania L, którą przyjmuje się zazwyczaj między pięć a dziesięć, oraz parametr α_p, który w praktyce ustala się na 0.5, zgodnie z sugestią zawartą w [44].

4.3 Filtr cząsteczkowy uwzględniający niskowymiarową rozmaitość

Odnosząc się do dotychczas opisanych metod, tj. zwykłego *filtra cząsteczkowego* i *wyżarzanego filtra cząsteczkowego*, można wyróżnić dwie pożądane cechy, które powinien posiadać *filtr cząsteczkowy*:

1. Ze względu na wysokowymiarowy charakter przestrzeni stanów, cząsteczki powinny koncentrować się w miejscach, gdzie prawdopodobieństwo a posteriori $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ jest

wysokie, w przeciwnym razie jakość uzyskanej estymaty $\hat{\mathbf{x}}_t$ będzie niska. Tę własność posiada jedynie *wyżarzany filtr cząsteczkowy*, które skupia cząsteczki wokół ekstremów lokalnych funkcji gęstości. Zwykły *filtr cząsteczkowy* w dużym stopniu pokrywa obszary o niskim prawdopodobieństwie.

2. Ze względu na wielomodalny charakter rozkładu $p(\mathbf{x}_t|\mathcal{I}_{1:t})$, cząsteczki powinny koncentrować się wokół wielu ekstremów lokalnych funkcji gęstości, w przeciwnym razie istnieje ryzyko, że system śledzący się zgubi. Tej własności zdecydowanie nie posiada *wyżarzany filtr cząsteczkowy*, w którym cząsteczki skupiają się zazwyczaj wokół jednego, a co najwyżej kilku dominujących ekstremów. Zwykły *filtr cząsteczkowy* posiada tę cechę jedynie, gdy liczba cząsteczek jest duża.

Biorąc pod uwagę spostrzeżenie opisane w rozdziale 3.2.1, że rzeczywiste trajektorie ruchu człowieka usytuowane są w pobliżu niskowymiarowej *rozmaitości*, można przypuszczać, że wartość funkcji gęstości będzie wysoka w jej pobliżu. Dodatkowo większość ekstremów odpowiadająca rzeczywistym konfiguracjom będzie również skoncentrowana w jej sąsiedztwie. Ponadto ze względu na fakt, że wymiar *rozmaitości* jest istotnie niższy od wymiaru przestrzeni stanów, liczba cząsteczek potrzebnych do pokrycia obszaru *rozmaitości* jest dużo mniejsza niż liczba potrzebna do pokrycia fragmentu przestrzeni stanów. W konsekwencji skupiając cząsteczki wokół *rozmaitości* zagwarantujemy, że po pierwsze pokryty będzie obszar o wysokim prawdopodobieństwie, a po drugie cząsteczki będą rozrzucone pomiędzy różnymi ekstremami lokalnymi. W tym celu został zaproponowany *filtr cząsteczkowy* rozwiązujący problem *filtrowania* (3.24), który uwzględnia wiedzę o niskowymiarowej *rozmaitości*.

Rozważmy najpierw rozkład a priori w chwili t, czyli przed uwzględnieniem bieżącej obserwacji $\mathcal{I}_t$:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t-1}) = \iint p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t)p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{z}_t. \tag{4.23}$$

Gdyby było możliwe wygenerowanie z niego próby, wtedy rozkład a posteriori mógłby być przybliżony zgodnie z zależnością (4.14). Niestety najczęściej postać rozkładu $p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t)$ nie pozwala wprost wygenerować realizacji $\mathbf{x}_t$ przy znajomości $\mathbf{x}_{t-1}$ i $\mathbf{z}_t$ i wymaga zastosowania rozkładu pomocniczego $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, o którym zakładamy, że zależy jedynie od $\mathbf{x}_{t-1}$.

Rozważmy następujący rozkład łączny:

$$\begin{aligned} p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) &= p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t, \mathcal{I}_{1:t-1})p(\mathbf{z}_t|\mathbf{x}_{t-1}, \mathcal{I}_{1:t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \\ &= p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \\ &= \frac{1}{Z}\tilde{p}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}), \end{aligned} \tag{4.24}$$

gdzie korzystamy z warunkowych niezależności wynikających z *probabilistycznego modelu grafowego* przedstawionego na rysunku 3.3 oraz zakładamy, że rozkład $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ występuje w postaci nieunormowanej, gdzie:

$$Z = \iint \tilde{p}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1:t}\mathrm{d}\mathbf{z}_t. \tag{4.25}$$

Wykorzystując ponadto rozkład pomocniczy możemy przekształcić rozkład łączny (4.24) w następujący sposób:

$$\begin{aligned} p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) &= \frac{1}{Z}\frac{\tilde{p}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}q(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \\ &= \frac{1}{Z}\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}), \end{aligned} \tag{4.26}$$

gdzie zostały wprowadzone współczynniki wagowe:

$$\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \frac{\tilde{p}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \tag{4.27}$$

oraz pomocniczy rozkład łączny:

$$q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}). \tag{4.28}$$

Następnie zakładając, że dysponujemy próbą $\overline{\mathcal{X}}_{t-1}$ z rozkładu a posteriori $p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$, możemy wygenerować próbę z pomocniczego rozkładu łącznego (4.28) korzystając z następującego schematu:

1. Dla ustalonej realizacji $\overline{\mathbf{x}}_{t-1}^{(n)}$ wygeneruj $\mathbf{z}_t^{(n)} \sim p(\mathbf{z}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$.
2. Dla ustalonej realizacji $\overline{\mathbf{x}}_{t-1}^{(n)}$ wygeneruj $\mathbf{x}_t^{(n)} \sim q(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$.

W ten sposób otrzymujemy zbiór realizacji $\left\{(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\right\}_{n=1}^{N}$ z rozkładu (4.28). W konsekwencji możemy zastosować aproksymację analogiczną do (4.2):

$$\hat{q}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = \frac{1}{N}\sum_{n=1}^{N}\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)})\delta(\mathbf{x}_{t-1} - \overline{\mathbf{x}}_{t-1}^{(n)})\delta(\mathbf{z}_t - \mathbf{z}_t^{(n)}), \tag{4.29}$$

gdzie rzeczywisty rozkład przybliżamy dyskretnym rozkładem skoncentrowanym w skończonej liczbie punktów. Następnie zastępując w (4.26) rozkład pomocniczy przez aproksymację (4.29) otrzymujemy:

$$\begin{aligned}\hat{p}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) &= \frac{1}{Z}\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)\hat{q}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{Z}\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)})\delta(\mathbf{x}_{t-1} - \overline{\mathbf{x}}_{t-1}^{(n)})\delta(\mathbf{z}_t - \mathbf{z}_t^{(n)}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{Z}\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)})\delta(\mathbf{x}_{t-1} - \overline{\mathbf{x}}_{t-1}^{(n)})\delta(\mathbf{z}_t - \mathbf{z}_t^{(n)}).\end{aligned} \tag{4.30}$$

Całkując powyższą zależność po zmiennych $\mathbf{z}_t$ i $\mathbf{x}_{t-1}$, dostajemy przybliżenie rozkładu a priori w chwili t:

$$\begin{aligned}\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t-1}) &= \frac{1}{Z}\iint \hat{p}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{z}_t\mathrm{d}\mathbf{x}_{t-1} \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{Z}\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}),\end{aligned} \tag{4.31}$$

co rozwiązuje krok predykcji (ang. prediction step) w zadaniu *filtrowania* (3.24). Powyższą postać możemy użyć do aproksymacji rozkładu a posteriori, uwzględniającego bieżącą obserwację $\mathcal{I}_t$, poprzez zastosowanie (4.31) do wyrażenia (4.11):

$$\begin{aligned}p(\mathbf{x}_t|\mathcal{I}_{1:t}) &\approx \frac{p(\mathcal{I}_t|\mathbf{x}_t)\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t-1})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \\ &= \frac{1}{N}\frac{1}{Z}\sum_{n=1}^{N}\frac{\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})}\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}).\end{aligned} \tag{4.32}$$

Stosując z kolei przybliżenie (4.29) do czynnika normującego $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$ otrzymujemy:

$$\begin{aligned}p(\mathcal{I}_t|\mathcal{I}_{1:t-1}) &\approx \int p(\mathcal{I}_t|\mathbf{x}_t)\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{x}_t \\ &= \frac{1}{N}\frac{1}{Z}\sum_{n=1}^{N}\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}).\end{aligned} \tag{4.33}$$

Następnie łącząc razem (4.32) i (4.33) dostajemy ostateczną postać aproksymacji dla rozkładu a posteriori:

$$\hat{p}(\mathbf{x}_t|\mathcal{I}_{1:t}) = \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{4.34}$$

gdzie unormowane parametry wagowe zostały zdefiniowane następująco:

$$\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})}{\sum_{j=1}^{N} \tilde{\pi}(\mathbf{x}_t^{(j)})\tilde{\omega}(\mathbf{x}_t^{(j)}, \overline{\mathbf{x}}_{t-1}^{(j)}, \mathbf{z}_t^{(j)})}. \tag{4.35}$$

W ten sposób rozwiązany został krok aktualizacji (ang. update step) dla zadania filtrowania (3.24). Warto zauważyć, że rozkład (4.34) jest w analogicznej postaci do (4.14), z tą różnicą, że wagi wyliczane są w inny sposób. Próba z rozkładu a posteriori $\overline{\mathcal{X}}_t$, podobnie jak w przypadku zwykłego *filtra cząsteczkowego*, może być wygenerowana z użyciem *ponownego próbkowania* z rozkładu (4.34).

Algorithm 3: Filtr cząsteczkowy do śledzenia ruchu człowieka uwzględniający strukturę niskowymiarowej rozmaitości

Wejście: Stan początkowy $\mathbf{x}_0$, sekwencja pomiarów $\mathcal{I}_{1:T}$

Wyjście: Sekwencja estymat wektora stanu $\hat{\mathbf{x}}_{1:T}$

Powiel początkowy stan $\mathbf{x}_0$ do zbioru $\overline{\mathcal{X}}_0 = \{\overline{\mathbf{x}}_0^{(1)}, \ldots, \overline{\mathbf{x}}_0^{(N)}\}$;

for $t = 1 : T$ **do**

- Wygeneruj próbę $\mathcal{Z}_t = \{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(N)}\}$ z modelu dynamiki rzutującego na niskowymiarową rozmaitość, gdzie $\mathbf{z}_t^{(n)} \sim p(\mathbf{z}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
- Wygeneruj próbę $\mathcal{X}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)}\}$ z pomocniczego modelu dynamiki, gdzie $\mathbf{x}_t^{(n)} \sim q(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
- Wylicz wartości $\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ korzystając z zależności (4.27);
- Wylicz wartości $\tilde{\pi}(\mathbf{x}_t^{(n)})$ korzystając z modelu wiarygodności $p(\mathcal{I}_t|\mathbf{x}_t)$;
- Znormalizuj uzyskane wartości do $\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ korzystając z (4.35);
- Wyznacz estymatę wektora stanu $\hat{\mathbf{x}}_t$ używając reguły (4.36) lub (4.37);
- Wygeneruj próbę $\overline{\mathcal{X}}_t = \{\overline{\mathbf{x}}_t^{(1)}, \ldots, \overline{\mathbf{x}}_t^{(N)}\}$ z rozkładu (4.34);

end

Reguły decyzyjne, na podstawie których wyznaczana jest estymata wektora stanu, mają analogiczną postać do (4.9) lub (4.10), tj.

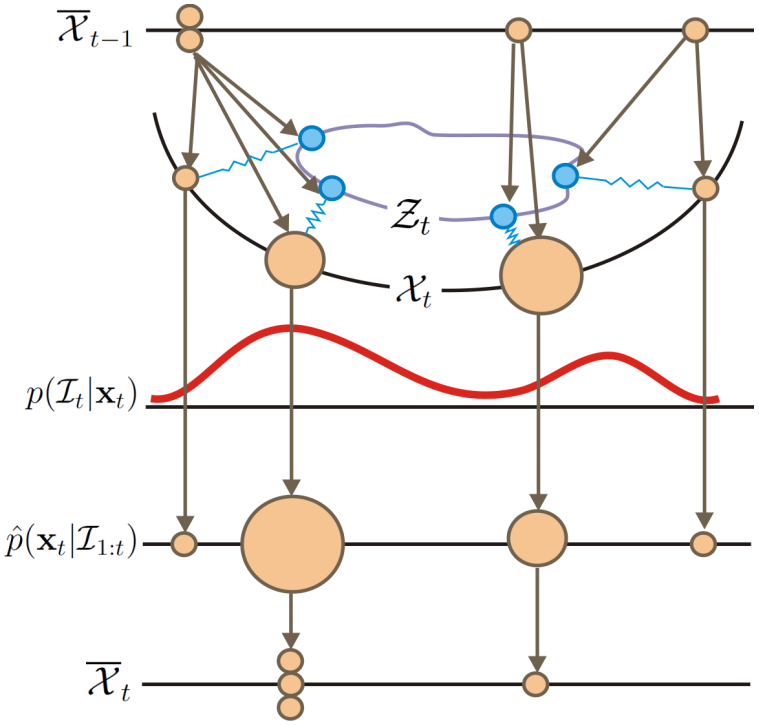
$$\hat{\mathbf{x}}_t = \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) \mathbf{x}_t^{(n)} \tag{4.36}$$

dla średniej z rozkładu lub alternatywnie

$$\hat{\mathbf{x}}_t = \arg\max_n \ \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) \tag{4.37}$$

dla estymatora maksymalnego a posteriori.

Podsumowując powyższe rozważania, został zaproponowany *filtr cząsteczkowy* uwzględniający strukturę niskowymiarowej *rozmaitości*, który rozwiązuje zadanie *filtrowania* (3.24). Procedura ta została opisana algorytmem 3.



Rysunek 4.2: Schemat działania filtra cząsteczkowego uwzględniającego strukturę niskowymiarowej rozmaitości.

Na rysunku 4.2 został przedstawiony schemat działania zaproponowanego algorytmu. Warto zwrócić uwagę, że zasadnicza różnica w stosunku do zwykłego *filtra cząsteczkowego*

(rysunek 4.1) polega na wygenerowaniu dodatkowego zbioru realizacji $\mathcal{Z}_t$, który zawiera informację o przewidywanym położeniu cząsteczek w układzie współrzędnych związanym z niskowymiarową *rozmaitością*. Ta informacja jest brana pod uwagę do ustalenia wag $\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ dla cząsteczek ze zbioru $\mathcal{X}_t$ wygenerowanego przy użyciu pomocniczego rozkładu $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. Intuicyjnie oznacza to, że im cząsteczki bardziej będą się oddalać od *rozmaitości*, tym przyznawane im wagi będą niższe.

Podsumowując, warto sformułować następujące uwagi dotyczące zaproponowanej procedury:

1. Wprowadzenie struktury niskowymiarowej *rozmaitości* zapewnia lepsze modelowanie rozkładu a posteriori $p(\mathbf{x}_t|\mathcal{I}_{1:t})$. W szczególności dodatkowa zmienna $\mathbf{z}_t$ o niższym wymiarze niż $\mathbf{x}_t$ ogranicza klasę wszystkich modeli a posteriori, które mogą być przybliżone przez aproksymację przy pomocy zbioru cząsteczek, wymuszając ich rozkład w pobliżu *rozmaitości*. Może to być postrzegane jako technika *regularyzacji* (ang. regularization) dla *filtra cząsteczkowego*, poprzez uwzględnienie apriorycznej wiedzy o niskowymiarowej strukturze.

2. Modele $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ i $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ wymagają uwzględnienia transformacji pomiędzy *rozmaitością* i przestrzenią stanów, które mają skomplikowany nieliniowy charakter, a ich wyznaczenie wymaga użycia zaawansowanych technik *redukcji wymiarów* (ang. dimensionality reduction).

Rozdział 5

Modele wiarygodności

W rozdziale zostały opisane trzy modele *modele wiarygodności* (ang. likelihood model). Dwa powszechnie stosowane w literaturze – model oparty na *sylwetkach* (ang. silhouette-based) i model oparty na krawędziach (ang. edge-based) oraz trzeci autorski – model oparty na lokalnych deskryptorach.

5.1 Modelowanie ciała

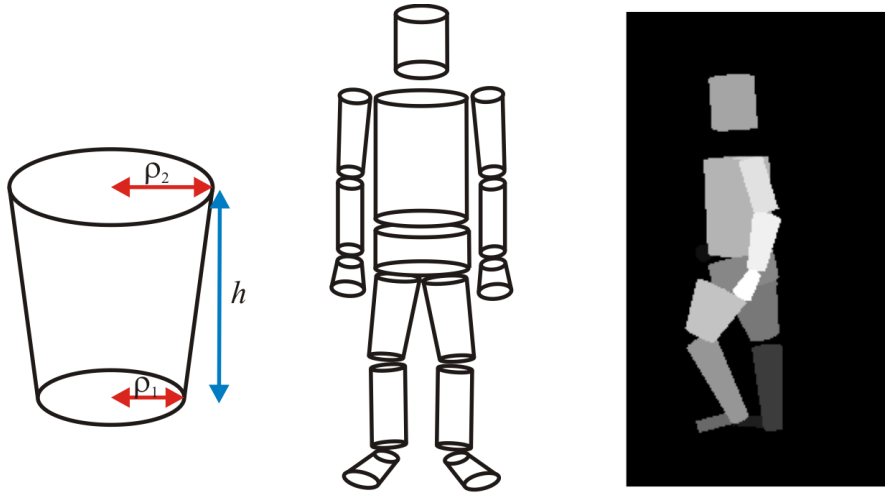
Wspólnym dla wszystkich *modeli wiarygodności* jest zaproponowanie *modelu ciała* (ang. body model), który będzie używany do wygenerowania wyglądu danej konfiguracji na podstawie ustalonego stanu $\mathbf{x}$. Innymi słowy, należy ustalić konkretną postać poszczególnych elementów sztywnych $\mathcal{V}$, wchodzących w skład *drzewa kinematycznego* opisanego w rozdziale 2.1.2.

W pracy użyto modelu opartego na ściętych stożkach, w którym każdy element sztywny jest postaci:

$$\mathcal{V} = \left\{ \mathbf{v} \in \mathbb{R}^3 : \; 0 \leqslant v_z \leqslant h, \; \sqrt{v_x^2 + v_y^2} \leqslant \frac{v_z}{h}\rho_2 + \left(1 - \frac{v_z}{h}\right)\rho_1 \right\}, \tag{5.1}$$

gdzie h oznacza wysokość ściętego stożka, a ρ_1 i ρ_2 oznaczają odpowiednio dolny i górny promień. Ścięty stożek został przedstawiony na rysunku 5.1a. Warto zauważyć, że jeśli oba promienie są równe, wówczas ścięty stożek staje się walcem.

Zbiór elementów sztywnych w postaci (5.1) połączonych poprzez strukturę *drzewa kinematycznego* nazywamy *modelem ciała* i oznaczać będziemy symbolem $\mathcal{B}$. Został on przed-



Rysunek 5.1: Modelowanie ciała człowieka. (a) Element sztywny w postaci ściętego stożka. (b) Model ciała. (c) Mapa głębi.

stawiony na rysunku 5.1b. Ze względu na swoją prostotę, jest to najczęściej stosowany w literaturze model [9, 32, 41, 44, 63, 86, 97, 114, 119, 120, 139, 140, 141, 153]. Jednakże powszechnie wykorzystuje się również inne, m.in. *modele ciała* oparte na elipsoidach [26, 34, 71, 109], kwadrykach [84, 144, 150, 151], konturach [56], trójwymiarowych siatkach [58, 59].

Użyty w pracy model (rysunek 5.1b) składa się z piętnastu elementów: miednica, tułów, głowa, uda, łydki, stopy, ramiona, przedramiona i dłonie. W literaturze spotka się inne kombinacje użytych części ciała, od mniej szczegółowych modeli ograniczających się tylko do górnej lub dolnej części człowieka, do bardziej szczegółowych modelujących na przykład poszczególne palce na dłoniach.

5.1.1 Mapa głębi

W przypadku rzutowania *modelu ciała* $\mathcal{B}$ na ustalony obraz, widziany z danej kamery, często konieczne jest stwierdzenie, które części ciała są z danej perspektywy widoczne i w jakim stopniu. W tym celu tworzy się *mapę głębi* (ang. depth map) D^{I}, która jest obrazem

o takich samych wymiarach jak obraz I, tj.

$$\mathrm{D}^{\mathrm{I}} = [\mathrm{D}^{\mathrm{I}}_{ij}], \tag{5.2}$$

gdzie wartości poszczególnych komórek oznaczają numer elementu sztywnego widocznego na danym pikselu. W przypadku, gdy nie jest widoczny żaden element, ustawiana jest inna wartość odpowiadająca za tło.

Niech $\mathcal{J}^{\mathrm{I}}(\mathcal{V})$ oznacza zbiór indeksów (i, j) pikseli obrazu I, na których widoczny jest element $\mathcal{V}$ po zrzutowaniu z użyciem zależności (2.46). Formalnie zbiór ten ma następującą postać:

$$\mathcal{J}^{\mathrm{I}}(\mathcal{V}) = \Big\{(i, j) : \ (i, j) \in \mathcal{P}_{\mathrm{I}}(\mathcal{V})\Big\}. \tag{5.3}$$

Należy zauważyć, że może on być bardzo efektywnie wyznaczony rzutując jedynie wybrane punkty z elementu sztywnego z użyciem zależności (2.45) oraz stosując algorytm do wypełniania wielokątów.

Dodatkowo przez $\mathbf{c}_i$ oznaczmy środek i-tego elementu w globalnym układzie współrzędnych. Może on być wyznaczony poprzez wyznaczenie położenia w globalnym układzie współrzędnych punktu $(0, 0, h_i/2)$ z elementu $\mathcal{V}_i$.

Algorithm 4: Mapa głębi

Wejście: Parametry kalibracyjne kamery $\mathbf{A}$, $\mathbf{R}_{\mathrm{I}}$, $\mathbf{u}_{\mathrm{I}}$, model ciała $\mathcal{B}$, środki $\mathbf{c}_0, \ldots, \mathbf{c}_K$

Wyjście: Mapa głębi D^{I}

1 Zainicjalizuj mapę głębi D^{I} wartościami oznaczającymi tło;
2 **for** $i = 0 : K$ **do**
3 | Wyznacz odległość kamery o środka elementu sztywnego, $d^{\mathrm{I}}_i := \|\mathbf{c}_i - \mathbf{u}_{\mathrm{I}}\|$;
4 **end**
5 Posortuj zbiór indeksów $\{0, \ldots, K\}$ malejąco według odległości $\{d^{\mathrm{I}}_0, \ldots, d^{\mathrm{I}}_K\}$ do zbioru indeksów $\{\xi(0), \ldots, \xi(K)\}$;
6 **for** $i = 0 : K$ **do**
7 | Wyznacz zbiór $\mathcal{J} := \mathcal{J}^{\mathrm{I}}(\mathcal{V}_{\xi(i)})$ zgodnie z zależnością (5.3);
8 | Wypełnij $\mathrm{D}^{\mathrm{I}}(\mathcal{J}) := \xi(i)$;
9 **end**

W celu wyznaczenia mapy głębi została zaproponowana heurystyka opisana algorytmem 4, gdzie $\mathrm{D}^{\mathrm{I}}(\mathcal{J})$ oznacza odwołanie do wszystkich pikseli ze zbioru $\mathcal{J}$. Należy zwrócić uwagę, że zaproponowana procedura nie zawsze prawidłowo oznacza widoczny w danym miejscu element. Wynika to z faktu, że odległość elementu sztywnego od kamery szacowana jest jedynie na podstawie odległości jego środka. Niemniej, jest ona bardzo efektywna pod względem obliczeniowym i wystarczająca na potrzeby dalszych metod. Efekt działania algorytmu został zaprezentowany na rysunku 5.1c, gdzie każdy element im jest jaśniejszy, tym bliżej kamery jest on położony.

5.2 Model oparty na sylwetkach

Idea modelu wiarygodności opartego na *sylwetkach* (ang. silhouette) polega na tym, aby na podstawie *wektora stanu* stworzyć binarny obraz, przedstawiający *sylwetkę modelu ciała* i porównać ją z *sylwetką* człowieka uzyskaną z rzeczywistego obrazu I.

W tym celu korzystając z definicji (5.3), wprowadzamy oznaczenie na zbiór pikseli, na których widoczny jest *model ciała*, tj. widoczny jest którykolwiek z elementów sztywnych:

$$\mathcal{J}^{\mathrm{I}}(\mathbf{x}) = \bigcup_{\mathcal{V}\in\mathcal{B}} \mathcal{J}^{\mathrm{I}}(\mathcal{V}). \tag{5.4}$$

Wtedy binarny obraz, na którym jest *sylwetka modelu ciała*, ma wartość jeden dla pikseli z powyższego zbioru oraz zero dla pozostałych. Zauważmy, że zbiór (5.4) został zdefiniowany jako funkcja *wektora stanu* $\mathbf{x}$. Wynika to z faktu, że każdy element sztywny rzutowany jest zgodnie z zależnością (2.46), która wymaga by współrzędne należących do niego punktów były wyrażone w globalnym układzie współrzędnych. Z kolei przejście od lokalnych do globalnych współrzędnych jest możliwe poprzez zastosowanie odpowiednich macierzy rotacji i wektorów przesunięć, które są zależne od bieżącego *wektora stanu*. Procedura ta została opisana w rozdziale 2.

5.2.1 Problem oddzielania tła

Do porównania *sylwetki* wygenerowanej z *modelu ciała* z rzeczywistym obrazem, potrzeba procedury, która wyekstrahuje z rzeczywistego obrazu analogiczną binarną *sylwetkę*.

Wymaga to określenia, które piksele na obrazie należą do śledzonego człowieka, a które należą do tła i prowadzi do znanego problemu *oddzielania tła* (ang. background subtraction).

Zagadnienie *oddzielania tła* jest jednym z fundamentalnych problemów w obszarze *widzenia komputerowego*, gdyż pozwala na odfiltrowanie nadmiarowej informacji (tła) zawartej w obrazie, która nie jest istotna w rozważanym zagadnieniu i traktowana jest jako szum. Skuteczne oddzielenie tła jest bardzo trudnym problemem, szczególnie w przypadkach, gdy ma ono charakter niestacjonarny, wynikający z występujących w nim ruchomych obiektów, zmiany warunków oświetlenia, przemieszczenia się kamery itp. Powstało wiele technik do radzenia sobie z tym problemem [121], w tym m.in. algorytmy adaptacyjne oparte na mieszaninie rozkładów prawdopodobieństwa [145], metody bazujące na modelach nieparametrycznych [49], techniki stosujące teorię *skompresowanego zrozumienia* (ang. compressive sensing) [31].

W pracy została zastosowana metoda oparta na prostym teście statystycznym. Algorytm ten zakłada, że każdy piksel należący do tła, może być scharakteryzowany łącznym rozkładem prawdopodobieństwie $p(\mathrm{I}_{ij}^R, \mathrm{I}_{ij}^G, \mathrm{I}_{ij}^B)$ na trzy kanały – czerwony, zielony i niebieski. Losowy charakter każdego piksela wynika przede wszystkim z szumów kamery, wahania oświetlenia i efektu skalowania zakresu każdego z kanałów, charakterystycznego dla kamer cyfrowych. Reguła, na bazie której podejmowana jest decyzja czy dany piksel należy do człowieka czy do tła, jest typowa dla problemów wykrywania anomalii, tj.

$$p(\mathrm{I}_{ij}^R, \mathrm{I}_{ij}^G, \mathrm{I}_{ij}^B) < p_0, \tag{5.5}$$

gdzie p_0 jest pewnym ustalonym progiem. Oznacza to, że jeśli wartość funkcji gęstości dla danego piksela (i, j) jest poniżej ustalonego progu, wtedy przyjmuje się, że piksel jest obserwacją anomalną i należy do człowieka.

Rozważane w pracy zbiory danych posiadają statyczne tło, dlatego możemy przyjąć, że każdy piksel charakteryzuje się rozkładem normalnym, niezależnie na każdym z kanałów:

$$p(\mathrm{I}_{ij}^c) = \mathcal{N}(\mathrm{I}_{ij}^c | \mu_{ij}^c, \sigma_{ij}^c), \tag{5.6}$$

gdzie $c \in \{R, G, B\}$, a parametrami rozkładu są wartość średnia i odchylenie standardowe. Wtedy rozkład łączny jest postaci:

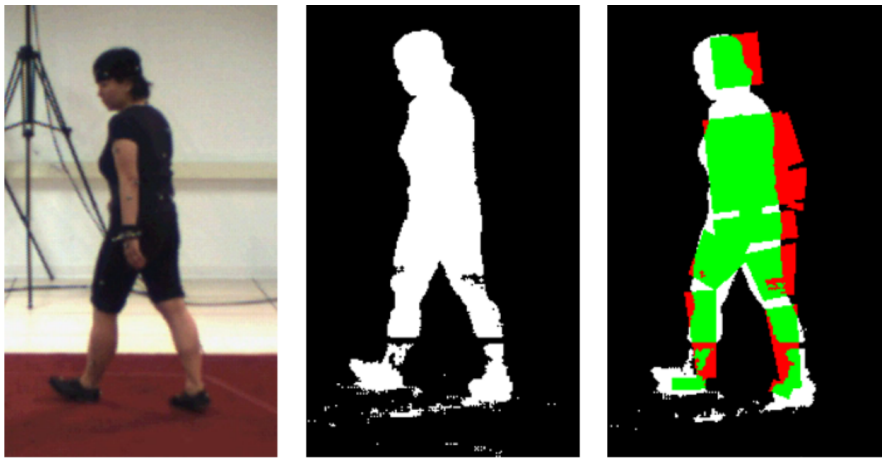
$$p(\mathrm{I}_{ij}^R, \mathrm{I}_{ij}^G, \mathrm{I}_{ij}^B) = p(\mathrm{I}_{ij}^R) p(\mathrm{I}_{ij}^G) p(\mathrm{I}_{ij}^B). \tag{5.7}$$

Dysponując zbiorem obserwacji samego tła $\mathcal{I}_{BG}$ możemy oszacować parametry rozkładu (5.6), korzystając z estymatorów największej wiarygodności dla rozkładu normalnego:

$$\mu_{ij}^{c} = \frac{1}{|\mathcal{I}_{BG}|} \sum_{\mathrm{I} \in \mathcal{I}_{BG}} \mathrm{I}_{ij}^{c}, \tag{5.8}$$

$$\sigma_{ij}^{c} = \sqrt{\frac{1}{|\mathcal{I}_{BG}|} \sum_{\mathrm{I} \in \mathcal{I}_{BG}} \left(\mathrm{I}_{ij}^{c} - \mu_{ij}^{c}\right)^{2}}, \tag{5.9}$$

gdzie $|\mathcal{I}_{BG}|$ oznacza liczność zbioru. W praktyce powyższe statystyki muszą być wyznaczone przed rozpoczęciem procesu śledzenia człowieka. Stosowany w pracy zbiór danych *HumanEva* [140] zawiera odpowiednie sekwencje $\mathcal{I}_{BG}$ dla każdej z kamer.



(a) (b) (c)

Rysunek 5.2: Model wiarygodności oparty na sylwetkach. (a) Wejściowy obraz I. (b) Binarna sylwetka S^{I}. (c) Porównanie sylwetki z obrazu wejściowego z sylwetką ze zrzutowanego modelu ciała.

Należy zatem stwierdzić, że korzystając z przedstawionej techniki *oddzielania tła*, możemy stworzyć binarny obraz $\mathrm{S}^{\mathrm{I}} = [\mathrm{S}_{ij}^{\mathrm{I}}]$ zawierający *sylwetkę* człowieka widocznego na obrazie I:

$$\mathrm{S}_{ij}^{\mathrm{I}} = \begin{cases} 1, & p(\mathrm{I}_{ij}^{R}, \mathrm{I}_{ij}^{G}, \mathrm{I}_{ij}^{B}) < p_0 \\ 0, & p(\mathrm{I}_{ij}^{R}, \mathrm{I}_{ij}^{G}, \mathrm{I}_{ij}^{B}) \geqslant p_0 \end{cases}. \tag{5.10}$$

Przykładowa binarna *sylwetka* została przedstawiona na rysunku 5.2b.

5.2.2 Funkcja wiarygodności

Model wiarygodności $p(\mathcal{I}|\mathbf{x})$ ma za zadanie ocenić na ile prawdopodobne jest, że na danej serii obrazów $\mathcal{I}$ widoczna jest konfiguracja $\mathbf{x}$. W tym celu porównane zostaną obrazy uzyskane na podstawie rzutowania *modelu ciała* (5.4) z obrazem powstałym w wyniku *oddzielania tła* (5.10). Prowadzi to do następującej postaci modelu:

$$-\ln p(\mathcal{I}|\mathbf{x}) = \frac{1}{|\mathcal{I}|}\sum_{\mathrm{I}\in\mathcal{I}}\left\{\frac{1}{|\mathcal{J}^{\mathrm{I}}(\mathbf{x})|}\sum_{(i,j)\in\mathcal{J}^{\mathrm{I}}(\mathbf{x})}\left(1-\mathrm{S}^{\mathrm{I}}_{ij}\right)\right\}+\mathrm{const.} \tag{5.11}$$

Zauważmy, że im niższa wartość funkcji w powyższej postaci, tym wartość wiarygodności jest wyższa, a więc *wektor stanu* jest lepiej dopasowany do obrazu. Dodatkowo należy nadmienić, że powyższa zależność podana jest w zlogarytmowanej formie z dokładnością do pewnej stałej, która jest czynnikiem normującym dla rozkładu $p(\mathcal{I}|\mathbf{x})$, a jej wyznaczenie wymagałoby wysumowania po przestrzeni wszystkich możliwych zbiorów obrazów i nie jest możliwe do wykonania. Ponieważ jednak *model wiarygodności* jest używany w algorytmach 1, 2 i 3 do wyliczenia wartości $\pi(\mathbf{x})$, to zamiast wyznaczać $\tilde{\pi}(\mathbf{x}) = p(\mathcal{I}|\mathbf{x})$, możemy skorzystać z nieunormowanej postaci modelu i wyznaczyć $\tilde{\pi}(\mathbf{x}) = \tilde{p}(\mathcal{I}|\mathbf{x})$, gdyż stałe normujące skrócą się w procesie normalizacji (4.7).

Na rysunku 5.2c zostało przedstawione porównanie obrazu S^{I} z *sylwetką* wygenerowaną na podstawie hipotetycznej konfiguracji $\mathbf{x}$. Kolorem zielonym zostały zaznaczone te piksele ze zbioru $\mathcal{J}^{\mathrm{I}}(\mathbf{x})$, dla których $S^{\mathrm{I}}_{ij} = 1$, a kolorem czerwonym pozostałe piksele z tego zbioru.

Na koniec warto zwrócić uwagę na kilka istotnych kwestii związanych z *modelem wiarygodności* opartym na *sylwetkach*:

1. Pomimo stosunkowo prostej postaci, model oparty na *sylwetkach* daje dobre wyniki w procesie *śledzenia ruchu człowieka* i dlatego jest powszechnie stosowany w literaturze [32, 41, 44, 86, 97, 140, 153]. Ponadto same *sylwetki*, uzyskane w procesie *oddzielania tła*, mają szersze zastosowanie w zagadnieniach *estymacji pozy* i używa się je często jako przykłady wchodzące w skład ciągu treningowego dla *modeli dyskryminacyjnych* lub do uczenia łącznych *rozmaitości* dla obrazów i konfiguracji [33, 66, 77, 108].

2. Funkcja wiarygodności w postaci (5.11) ocenia jedynie na ile *sylwetka* wygenerowana z modelu ciała pokryła sylwetkę z obrazu. Zagadnienie to można odwrócić i ocenić,

na ile *sylwetka* z obrazu pokrywa *sylwetkę* z *modelu ciała*. W pracy [140] został zaproponowany model łączący te dwie informację ze sobą:

$$-\ln p(\mathcal{I}|\mathbf{x}) = \frac{1}{|\mathcal{I}|}\sum_{\mathrm{I}\in\mathcal{I}}\left\{\frac{1}{|\mathcal{J}^{\mathrm{I}}(\mathbf{x})|}\sum_{(i,j)\in\mathcal{J}^{\mathrm{I}}(\mathbf{x})}\left(1-\mathrm{S}^{\mathrm{I}}_{ij}\right)\right.$$
$$\left.+\frac{1}{|\{\mathrm{S}^{\mathrm{I}}_{ij}=1\}|}\sum_{(i,j)\in\{\mathrm{S}^{\mathrm{I}}_{ij}=1\}}\mathbb{1}\Big\{(i,j)\notin\mathcal{J}^{\mathrm{I}}(\mathbf{x})\Big\}\right\}+\mathrm{const}, \tag{5.12}$$

gdzie $\{\mathrm{S}^{\mathrm{I}}_{ij}=1\}$ oznacza zbiór tych pikseli (i,j), dla których $\mathrm{S}^{\mathrm{I}}_{ij}=1$, a $\mathbb{1}(\cdot)$ oznacza indykator, który przyjmuje wartość 1, jeśli argument w postaci warunku logicznego jest prawdziwy i 0 w przeciwnym wypadku. Wiarygodność w tej postaci często daje lepsze rezultaty, jednak charakteryzuje się również wyższą złożonością obliczeniową w stosunku do (5.11).

3. Ocena wiarygodności odbywa się na zasadzie porównywania binarnych obrazów. Tracona jest wówczas informacja o kolorze i teksturze, która jest szczególnie istotna przy odtwarzaniu konfiguracji kończyn górnych, gdyż przez większą część ruchu znajdują się one na tle tułowia. W konsekwencji *modele wiarygodności* oparte na binarnych *sylwetkach* źle sobie radzą z prawidłową estymacją ich konfiguracji.

4. Technika *oddzielania tła* zaproponowana w pracy zakłada, że każdy piksel traktowany jest niezależnie. Powoduje to, że czasami fragmenty ciała znajdują się na tle o bardzo zbliżonym kolorze i w konsekwencji są wycinane z *sylwetki*, tworząc w nich dziury. Dodatkowo jakość *sylwetek* obniża również fakt, że często cień rzucany przez człowieka także jest oddzielany od tła. W konsekwencji spada również skuteczność algorytmów *śledzenia ruchu*, gdyż wiarygodność nie jest wtedy prawidłowo oceniana. Jakość uzyskiwanych *sylwetek* może być poprawiana poprzez dodatkowe odszumienie obrazu, wypełnienie brakujących fragmentów itp.

5. Metody śledzenia oparte na *filtrach cząsteczkowych* wymagają wyliczenia funkcji (5.11) dla każdej cząsteczki $\mathbf{x}^{(n)}$, których zazwyczaj używa się od kilkuset do kilku tysięcy. Wiąże się to z istotnym czasem obliczeniowym i powoduje, że jednowątkowe implementacje tych metod nie nadają się do śledzenia w czasie rzeczywistym. Niemniej implementacja wielowątkowa (np. na procesor graficzny GPU) może być tutaj wykonana wprost, ze względu na fakt, że wiarygodność może być wyliczona niezależnie

dla każdej z cząsteczek. Dodatkowo sama procedura wyznaczania pojedynczej wiarygodności może być usprawniona poprzez segmentację *sylwetek* na większe spójne obszary [111].

5.3 Model oparty na krawędziach

Koncepcja *modelu wiarygodności* opartego na krawędziach (ang. edge) polega na tym, aby zliczyć odsetek punktów rozłożonych wzdłuż krawędzi *modelu ciała*, które trafiły w krawędzie części ciała wyekstrahowane z obrazu. Intuicyjnie, im więcej punktów trafi w krawędzie na obrazie, tym model lepiej jest dopasowany i wartość wiarygodności wyższa.

Wprowadźmy analogiczną definicję do (5.3) i oznaczmy przez $\mathcal{E}^{\mathrm{I}}(\mathcal{V})$ zbiór pikseli rozłożonych wzdłuż krawędzi elementu sztywnego zrzutowanego na obraz I. Oczywiście zbiór ten może być efektywnie wyznaczony rzutując jedynie odpowiednio wybrane punkty z elementu sztywnego, a następnie łącząc je prostymi. Następnie wprowadźmy zbiór wszystkich punktów rozłożonych wzdłuż krawędzi modelu ciała $\mathcal{B}$:

$$\mathcal{E}^{\mathrm{I}}(\mathbf{x}) = \bigcup_{\mathcal{V}\in\mathcal{B}} \mathcal{E}^{\mathrm{I}}(\mathcal{V}), \tag{5.13}$$

gdzie podobnie jak przypadku (5.4) zbiór ten może być wyrażony jako funkcja stanu $\mathbf{x}$.

Zauważmy, że nie wszystkie punkty z (5.13) są widoczne z ustalonej perspektywy. W szczególności, jeśli całość lub fragment części ciała będzie przesłonięty przez inną część ciała, wtedy dana krawędź nie będzie widoczna w całości lub w części. Korzystając z *mapy głębi* D^{I} wyznaczonej z użyciem algorytmu 4, możemy zdefiniować wskaźnik, który określa czy dany punkt jest widoczny:

$$\upsilon^{\mathrm{I}}_{ij} = \begin{cases} 1, & \exists_k \ (i,j) \in \mathcal{E}^{\mathrm{I}}(\mathcal{V}_k) \ \wedge \ \mathrm{D}^{\mathrm{I}}_{ij} = k \\ 0, & \text{w przeciwnym przypadku} \end{cases} . \tag{5.14}$$

Powyższy wskaźnik przyjmuje wartość jeden wtedy i tylko wtedy, gdy punkt (i,j) należy do krawędzi *modelu ciała* i jest widoczny, tj. indeks elementu sztywnego pokrywa się z wartością na *mapie głębi*.

5.3.1 Mapa krawędzi

Podobnie jak w przypadku modelu opartego na *sylwetkach*, obraz wejściowy należy przetransformować do takiej postaci, aby możliwe było porównanie go ze zbiorem punktów rozłożonych wzdłuż krawędzi *modelu ciała*. Innymi słowy, należy wyekstrahować krawędzie z wejściowego obrazu I. Prowadzi to do powszechnie znanego w dziedzinie *przetwarzania obrazów* (ang. image processing) problemu *detekcji krawędzi* (ang. edge detection) [55, 138], dla którego istnieje wiele skutecznych algorytmów, przykładowo *Canny Edge Detector* [27].

W pracy zastosowano autorską, uproszczoną procedurę wykrywania krawędzi. W pierwszej kolejności wejściowy kolorowy obraz I jest konwertowany do obrazu w skali szarości, stosując konwersję polegającą na liniowej kombinacji kolorów:

$$\mathrm{I}_{ij} = 0.299 \cdot \mathrm{I}^R_{ij} + 0.587 \cdot \mathrm{I}^G_{ij} + 0.114 \cdot \mathrm{I}^B_{ij}. \tag{5.15}$$

Następnie na obraz nakłada się pionowy i poziomy filtr gradientowy, o następujących postaciach:

$$\mathrm{F}^x = \begin{bmatrix} -1/2 & 0 & 1/2 \end{bmatrix}, \tag{5.16}$$

$$\mathrm{F}^y = \begin{bmatrix} 1/2 \\ 0 \\ -1/2 \end{bmatrix}. \tag{5.17}$$

Nałożenie filtra na obraz polega na wykonaniu operacji splotu (ang. convolution) pomiędzy obrazem i filtrem, którą definiuje się następująco [1] :

$$(\mathrm{I} * \mathrm{F})_{ij} = \sum_{m,n} \mathrm{I}_{mn} \mathrm{F}_{i-m,j-n}. \tag{5.18}$$

Wynikiem powyższego działania jest obraz o takich samych wymiarach, jak I. Intuicyjnie operacja splotu polega na lokalnym liczeniu iloczynu skalarnego pomiędzy obrazem i filtrem. Wartość iloczynu skalarnego jest tym wyższa, im bardziej obraz lokalnie podobny jest do filtra. Oznacza to, że filtr poziomy (5.16) daje silne odpowiedzi w miejscu, gdzie

[1] Domyślnie przyjmuje się konwencję, że środkowy element filtra ma indeks $(0, 0)$. Pozostałe elementy w zależności od ich położenia mają indeksy dodatnie lub ujemne.

występują pionowe krawędzie, natomiast filtr pionowy (5.17) w miejscu, gdzie krawędzie są poziome. Bazując na definicji splotu, wprowadźmy następujące oznaczenia:

$$\mathrm{G}^{\mathrm{I},x} = \mathrm{I} * \mathrm{F}^{x}, \tag{5.19}$$

$$\mathrm{G}^{\mathrm{I},y} = \mathrm{I} * \mathrm{F}^{y}. \tag{5.20}$$

Korzystając z uzyskanych obrazów, można stworzyć binarny obraz G^{I}, na którym zostaną wyróżnione krawędzie:

$$\mathrm{G}^{\mathrm{I}}_{ij} = \begin{cases} 1, & (G^{\mathrm{I},x}_{ij})^2 + (G^{\mathrm{I},y}_{ij})^2 > \eta^2 \\ 0, & (G^{\mathrm{I},x}_{ij})^2 + (G^{\mathrm{I},y}_{ij})^2 \leqslant \eta^2 \end{cases} . \tag{5.21}$$

Warto zwrócić uwagę, że para $(\mathrm{G}^{\mathrm{I},x}_{ij}, \mathrm{G}^{\mathrm{I},y}_{ij})$ oznacza wektor gradientu w punkcie (i, j). Zgodnie z definicją (5.21) krawędź zostanie wykryta na obrazie, jeśli długość wektora gradientu przekroczy pewien ustalony próg η.

Obraz w postaci (5.21) wystarczyłby do porównania z krawędziami uzyskanymi na podstawie *modelu ciała* (5.13). Zważywszy jednak na niedoskonałości *modelu ciała*, wynikające z uproszczenie budowy człowieka przy pomocy ściętych stożków, często zdarza się, że krawędzie uzyskane przy pomocy detektora krawędzi nie pokrywają się idealnie z krawędziami z *modelu ciała*, nawet jeśli model reprezentuje prawidłową konfigurację ciała. Efekt ten można zredukować poprzez „rozmycie" obrazu (5.21) w taki sposób, by wartości pikseli na obrazie G^{I} były dodatnie jeszcze w pewnym otoczeniu krawędzi i malały wraz z odległością. Wtedy wystarczy, aby krawędź z modelu znalazła się w pobliżu krawędzi z obrazu, by wiarygodność danej konfiguracji była wysoka. W celu uzyskania rozmycia można zastosować filtr gaussowski F^{g} [2] , którego elementy wyznacza się na podstawie zależności:

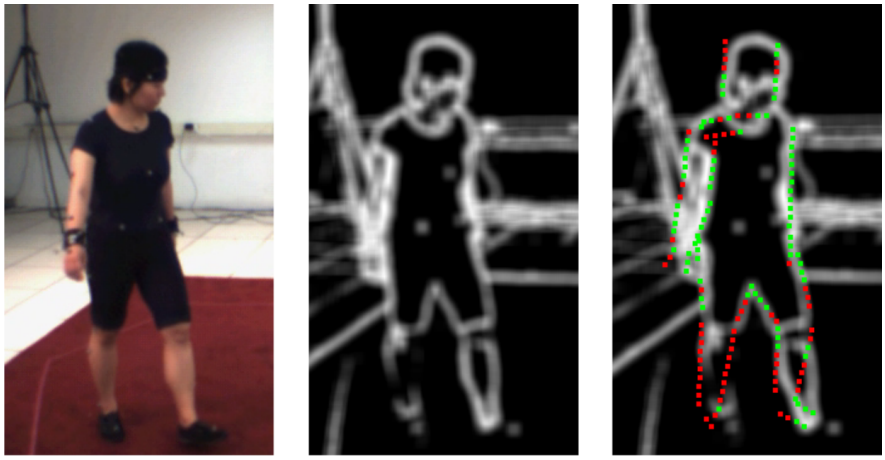
$$\mathrm{F}^{g}_{ij} = \frac{1}{2\pi\sigma_g^2} \exp\left(\frac{i^2 + j^2}{2\sigma_g^2}\right). \tag{5.22}$$

Ostateczną mapę krawędzi otrzymujemy stosując powyższy filtr do obrazu (5.21):

$$\mathrm{E}^{\mathrm{I}} = \mathrm{F}^{g} * \mathrm{G}^{\mathrm{I}}. \tag{5.23}$$

Uzyskany obraz może wymagać normalizacji tak, by wartości wszystkich pikseli zawierały się w przedziale $[0, 1]$. Na rysunku 5.3b została przedstawiona przykładowa mapa krawędzi uzyskana przy pomocy wyżej opisanej procedury.

[2] Rozmiar stosowanego filtra powinien zależeć o rozdzielczości obrazu, do którego się go stosuje. W pracy użyto filtry gaussowskie o rozmiarach 11×11.



(a) (b) (c)

Rysunek 5.3: Model wiarygodności oparty na krawędziach. (a) Wejściowy obraz I. (b) Mapa krawędzi E^{I}. (c) Porównanie mapy krawędzi z obrazu wejściowego z krawędziami ze zrzutowanego modelu ciała.

5.3.2 Funkcja wiarygodności

Do zdefiniowania *modelu wiarygodności* $p(\mathcal{I}|\mathbf{x})$ opartego na krawędziach korzystamy ze zbioru punktów rozłożonych wzdłuż krawędzi kończyn w *modelu ciała* (5.13), który należy porównać z mapą krawędzi (5.23) uzyskaną z wejściowego obrazu. Dodatkowo należy uwzględnić fakt, które punkty z krawędzi są widoczne z danej perspektywy, korzystając z zależności (5.14). Prowadzi to do następującej postaci modelu:

$$-\ln p(\mathcal{I}|\mathbf{x}) = \frac{1}{|\mathcal{I}|}\sum_{\mathrm{I}\in\mathcal{I}}\left\{\frac{1}{|\mathcal{E}^{\mathrm{I}}(\mathbf{x})|}\sum_{(i,j)\in\mathcal{E}^{\mathrm{I}}(\mathbf{x})}\left[\upsilon^{\mathrm{I}}_{ij}(1-\mathrm{E}^{\mathrm{I}}_{ij})+(1-\upsilon^{\mathrm{I}}_{ij})e_0\right]\right\}+\mathrm{const}, \tag{5.24}$$

gdzie $e_0 \in [0,1]$ oznacza stałą wartość, która jest przyznawana punktowi (i,j), jeśli jest on niewidoczny z danej perspektywy. Parametr ten ustawia się zazwyczaj na 0.5, faworyzując sytuacje, w których widocznych jest więcej punktów „trafiających" w krawędzie z mapy krawędzi. Podobnie, jak w przypadku wiarygodności opartej na *sylwetkach* (5.11), model zdefiniowany jest z dokładnością do stałej normującej, której wartość nie jest istotna do wyznaczenia wag $\pi(\mathbf{x})$.

Na rysunku 5.3c zostało przedstawione porównanie punktów ze zrzutowanego modelu z mapą krawędzi. Na zielono zostały zaznaczone widoczne punkty, dla których wartość $\mathrm{E}_{ij}^{\mathrm{I}}$ jest większa od $1 - e_0$, a na czerwono pozostałe widoczne punkty.

Na koniec należy podkreślić kilka charakterystycznych cech dla modelu wiarygodności opartego na krawędziach:

1. Model (5.24) stosuje się zazwyczaj jako formę pomocniczą dla modelu opartego na *sylwetkach* [32, 44, 140]. Ma to na celu dostarczenie dodatkowej informacji o wyglądzie człowieka, przykładowo o ułożeniu kończyn górnych. W większości przypadków stosowanie samego modelu opartego na krawędziach daje słabe rezultaty.

2. Model jest wrażliwy na wszelkie krawędzie, które nie pochodzą od człowieka. Przez to daje wysoką wiarygodność konfiguracjom, w których część kończyn dopasowuje się do elementów tła. Na rysunku 5.3a i 5.3b można zaobserwować silne krawędzie pochodzące od tła. Dlatego model dawał skuteczne rezultaty przede wszystkim w eksperymentach, gdzie tło było jednolite [44].

3. W przypadku, gdy kontrast między kończynami a tłem jest niski, krawędzie mogą nie zostać wyekstrahowane. Efekt ten można zauważyć w przypadku prawej nogi na rysunku 5.3b.

4. Kolejny problem wynika z dużej niedokładności *modelu ciała*. Przez to nawet poprawne konfiguracje kończyn, mogą nie dopasować się prawidłowo do obrazu. Efekt ten widać w przypadku głowy na rysunku 5.3c. Problem ten można zniwelować poprzez zastosowanie dokładnych *modeli ciała*, na przykład opartych na siatce [58, 59].

5.4 Model oparty na lokalnych deskryptorach

Niedoskonałości przedstawionych dotychczas *modeli wiarygodności* wynikają przede wszystkim z braku lub niejednoznaczności informacji na temat wyglądu człowieka, którą uzyskujemy z wejściowego obrazu. W szczególności w przypadku modelu opartego na *sylwetkach* często tracimy informację o położeniu kończyn górnych, gdyż zostają one nałożone na większy objętościowo tułów. Z drugiej strony model oparty na krawędziach wrażliwy jest

na nadmiarową informację pochodzącą z otoczenia, która może być mylnie interpretowana jako część ludzkiego ciała. Prowadzi to do konieczności stworzenia opisu wyglądu ciała, który będzie posiadał następujące dwie cechy:

1. Każda widoczna część ciała wnosi pewną informację, poprawiającą wartość funkcji wiarygodności.

2. Opis wyglądu poszczególnych części ciała jest możliwie unikatowy, tak by nie były one mylone z innymi elementami sceny.

W celu spełnienia pierwszego wymagania, na każdej części *modelu ciała* wyróżniono kilka punktów, z których każdy jest brany do wyliczenia funkcji wiarygodności. Oznaczmy przez $\mathcal{M}^{\mathrm{I}}(\mathcal{V})$ zbiór pikseli odpowiadający wyróżnionym punktom dla elementu $\mathcal{V}$, po rzutowaniu na obraz I. Elementami tego zbioru są trójki (i, j, k), gdzie (i, j) oznacza położenie piksela, a k unikatowy indeks wyróżnionego punktu. Jego postać jest różna od składowych zbioru (5.13), gdzie elementami były pary (i, j), ponieważ tam nie zakładaliśmy, że punkty rozłożone wzdłuż krawędzi są unikatowe. Następnie definiujemy zbiór wszystkich punktów wyróżnionych na *modelu ciała* po zrzutowaniu na obraz I:

$$\mathcal{M}^{\mathrm{I}}(\mathbf{x}) = \bigcup_{\mathcal{V} \in \mathcal{B}} \mathcal{M}^{\mathrm{I}}(\mathcal{V}). \tag{5.25}$$

Nawiązując do drugiego wymagania, dla każdego z wyróżnionych punktów zaproponowano pewien model, który charakteryzuje ich lokalny wygląd. Zbiór modeli dla wszystkich punktów składa się na *model wyglądu* (ang. appearance model), która stanowi uzupełnienie *modelu ciała* o dodatkową informację. Procedura tworzenia *modelu wyglądu* została podzielona na dwa etapy:

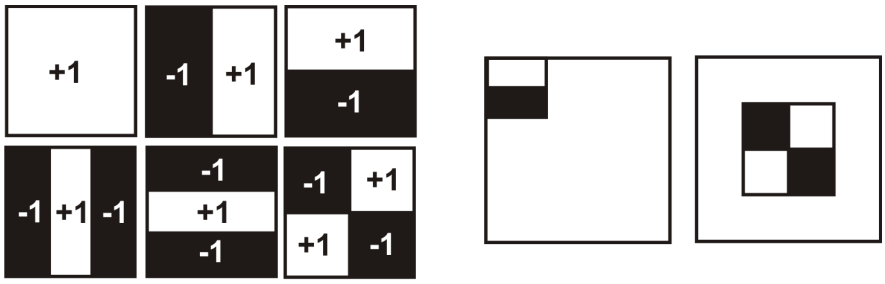
1. Pierwszy z nich wymaga określenia zestawu wartości $\boldsymbol{\phi} = (\phi^1, \ldots, \phi^M)$, który charakteryzuje wygląd ustalonego punktu na obrazie. Zestaw wyróżnionych w ten sposób cech nazywa się lokalnym deskryptorem (ang. local descriptor) dla punktu.

2. Drugi etap polega na nauczeniu modeli wchodzących w skład *modelu wyglądu* w taki sposób, aby określały w jakim stopniu lokalny deskryptor punktu na obrazie odpowiada wyglądowi punktu wyróżnionego na *modelu ciała*.

Ostateczna ocena wiarygodności polega na określeniu jak bardzo lokalne deskryptory wyliczone w punktach ze zbioru (5.25) są podobne do punktów scharakteryzowanych przez *model wyglądu*.

5.4.1 Lokalne deskryptory

Lokalne deskryptory powinny wnosić możliwie dużo informacji o otoczeniu wokół punktu, który opisują, jak kształt, kolor, tekstura itp. Dodatkowo ważne jest, aby były niewrażliwe na niewielkie przesunięcia obrazu, zmiany skali, rotacje i lokalne szumy. W pracy zastosowano deskryptory oparte na *falkach Haara* (ang. Haar wavelet), bazując na pomyśle zaproponowanym w [167]. Są one z powodzeniem wykorzystywane w wielu zagadnieniach *widzenia komputerowego*, w tym do opisu wyglądu części ciała [96, 142].



Rysunek 5.4: Filtry oparte na falkach Haara. (a) Rodzaje filtrów. (b) Przykładowe położenia filtra w otoczeniu punktu.

Ogólna idea tych deskryptorów polega na tym, aby zestaw cech, który opisuje otoczenie ustalonego punktu, wyrażony był jako odpowiedzi filtrów w postaci *falek Haara*. Niech $\{\mathrm{H}^1, \ldots, \mathrm{H}^M\}$ oznacza zbiór tych filtrów. Poszczególne elementy z tego zbioru tworzone są poprzez zmianę skali i przesunięcie sześciu typów filtrów przedstawionych na rysunku 5.4a. Dwa przykładowe filtry otrzymane w tym procesie zostały przedstawione na rysunku 5.4b, gdzie środek kwadratu, w którym są zawarte, jest punktem, dla którego wyliczany jest deskryptor. Odpowiedź filtra w punkcie (i, j) otrzymywana jest jako iloczyn skalarny filtra i obrazu:

$$\phi_{ij}^m = \sum_{k,l} \mathrm{I}_{kl} \mathrm{H}_{k-i,l-j}^m. \tag{5.26}$$

W ten sposób uzyskujemy deskryptor punktu $\boldsymbol{\phi}_{ij} = (\phi_{ij}^1, \ldots, \phi_{ij}^M)$. Należy zauważyć, że w przypadku obrazu kolorowego deskryptor wyliczany jest dla każdego kanału osobno, a ostateczny opis punktu tworzy złożenie tych trzech deskryptorów.

Warto zwrócić uwagę, że zazwyczaj wymiar wektora $\boldsymbol{\phi}_{ij}$ waha się od kilkuset do nawet kilku tysięcy składowych. To znaczy, że dla pojedynczego punktu ze zbioru (5.25), trzeba wyliczyć tyle odpowiedzi postaci (5.26). Procedura ta musi być powtórzona dla wszystkich wektorów stanu ze zbioru $\mathcal{X}_t$ w algorytmach przedstawionych w rozdziale 4, co w oczywisty sposób wiąże się z dużym nakładem obliczeniowym. Ponieważ jednak postać stosowanych filtrów składa się ze spójnych prostokątnych obszarów o stałej wartości -1 lub 1 (rysunek 5.4a), to poszczególne odpowiedzi mogą być wyliczone bardzo efektywnie z użyciem *obrazów całkowych* (ang. integral image) [167].

Obraz całkowy $\mathrm{II} = [\mathrm{II}_{ij}]$ jest wyliczany z wejściowego obrazu I poprzez scałkowanie go odpowiednio po obu składowych:

$$\mathrm{II}_{ij} = \sum_{k \leqslant i} \sum_{l \leqslant j} \mathrm{I}_{kl}. \tag{5.27}$$

Przy jego użyciu wyliczenie odpowiedzi dla jednego spójnego obszaru filtra wymaga wykonania jedynie czterech operacji arytmetycznych. Wynika to z faktu, że:

$$\begin{aligned} \sum_{i=a}^{b} \sum_{j=c}^{d} \mathrm{I}_{ij} \mathrm{H}_{k-i,l-j} &= h_{ca} \sum_{i=a}^{b} \sum_{j=c}^{d} \mathrm{I}_{ij} \\ &= h_{ca} \Big[\mathrm{II}_{bd} + \mathrm{II}_{a-1,c-1} - \mathrm{II}_{b,c-1} - \mathrm{II}_{a-1,d} \Big], \end{aligned} \tag{5.28}$$

gdzie $h_{ca} \in \{-1, 1\}$ oznacza wartość filtra na spójnym obszarze. Dzięki temu odpowiedzi stosowanych filtrów (rysunek 5.4a) mogą być otrzymane poprzez wykonanie od czterech do szesnastu operacji arytmetycznych, w zależności od liczby występujących w nich spójnych obszarów. Warto dodać, że technika *obrazów całkowych* miała przełomowe znaczenie w dziedzinie *widzenia komputerowego*, gdyż w sposób istotny zwiększyła wydajność wielu metod służących do detekcji i rozpoznawania obiektów na zdjęciach. Powstały również rozszerzenia tej procedury, m.in. histogramy całkowe [125], obrazy całkowe do liczenia macierzy kowariancji [160].

Na koniec należy podkreślić kilka istotnych kwestii dotyczących lokalnych deskryptorów:

1. *Falki Haara* są powszechnie stosowanym narzędziem w teorii kodowania sygnałów, gdyż w swojej podstawowej wersji tworzą ortogonalną bazę funkcji i pozwalają na wyrażenie sygnału jako kombinacja liniowa elementów z tej bazy. Należy jednak podkreślić, że stosowany w pracy zestaw filtrów $\{\mathrm{H}^1, \ldots, \mathrm{H}^M\}$ nie tworzy ortogonalnej bazy, a tym samym fragment obrazu wokół punktu (i, j) nie może być w prosty sposób rekonstruowany na podstawie deskryptora $\boldsymbol{\phi}_{ij}$.

2. W pracy założono, że okno wokół punktu, w którym wyliczane są filtry (rysunek 5.4b) ma stały rozmiar. To założenie ma sens jedynie wtedy, gdy śledzony człowiek znacząco nie oddala się ani nie przybliża względem obserwujących go kamer. W przeciwnym razie wielkości jego części ciała będą się istotnie zmieniać i nie będą mogły być opisane przez deskryptor o stałej wielkości okna. Stosuje się wtedy reprezentację przy pomocy tzw. piramidy w przestrzeni skal (ang. scale-space) [99], gdzie lokalny deskryptor wyliczany jest w wielu skalach jednocześnie. Innymi słowy, wyznaczany jest dla wielu wielkości okna, symbolizowanych przez różne poziomy piramidy.

3. Ponieważ konstrukcja *modelu wiarygodności* opartego na lokalnych deskryptorach jest niezależna od wyboru konkretnej postaci deskryptora, należy podkreślić, że istnieje wiele alternatywnych metod konstrukcji $\boldsymbol{\phi}$, które mogą z powodzeniem zastąpić *falki Haara*. Wymieniając kilka najważniejszych, można zastosować m.in. *Geometric blur* (GB) [14], *HMAX* [137], *Histogram of oriented gradients* (HOG) [39], *Hyperfeatures* [3], *Scale-invariant feature transform* (SIFT) [100].

4. Z względu na fakt, że postać filtrów $\{\mathrm{H}^1, \ldots, \mathrm{H}^M\}$ jest niezależna od obrazu, zazwyczaj większość z nich nie wnosi istotnej informacji do deskryptora $\boldsymbol{\phi}$. Można skonstruować zestaw filtrów (inny niż filtry Haara), który będzie zależny od specyfiki obrazów, dla których tworzone są deskryptory. Prowadzi to zwykle do bardziej informacyjnych filtrów. Takie zestawy można tworzyć przykładowo z użyciem metod *uczenia słowników* (ang. dictionary learning) [105] lub w wyniku uczenia *maszyn Boltzmanna* (ang. Boltzmann machine) [12].

5.4.2 Model wyglądu

Intuicyjnie model opisujący wygląd ustalonego punktu na ciele człowieka powinien charakteryzować się tym, że będzie reagował pozytywnie, jeśli punkt na obrazie będzie miał podobny wygląd. Okazuje się, że do skutecznego śledzenia konieczne jest również, aby model reagował silnie negatywnie, jeśli punkt na obrazie ma inny wygląd. Prowadzi to zagadnienia uczenia binarnych klasyfikatorów, gdzie parametry klasyfikatora ustala się tak, by możliwe było rozróżnienie przykładów należących do dwóch różnych klas.

W pracy wykorzystany został klasyfikator *Support Vector Machine* (SVM) [135], który obecnie jest standardowym narzędziem do klasyfikacji binarnej ze względu na swoją wysoką skuteczność, które może być formalnie uzasadniona na bazie *statystycznej teorii uczenia* [165, 166].

Załóżmy, że dysponujemy zbiorem deskryptorów opisujących wygląd ustalonego punktu na ciele w wielu przypadkach, tj. przy różnych konfiguracjach ciała, różnym położeniu na scenie, z innych kamer itp. Oczywiście deskryptory opisują otoczenia ustalonego punktu na dwuwymiarowym obrazie, a zatem dla wszystkich tych przypadków będą miały inną wartość. Dodatkowo załóżmy, że posiadamy zbiór deskryptorów nieopisujących wyróżnionego punktu na ciele. Przykładowo mogą to być opisy innych fragmentów ciała lub losowo wybranych fragmentów sceny. Zdefiniujmy następujący zbiór przykładów:

$$\mathcal{D}_\phi = \left\{(\boldsymbol{\phi}^n, y^n)\right\}_{n=1}^N, \tag{5.29}$$

gdzie $y^n \in \{-1, 1\}$ oznacza czy n-ty deskryptor opisuje wyróżniony punkt na ciele ($y_n = 1$) czy inny fragment obrazu ($y_n = -1$).

Wtedy problem uczenia klasyfikatora *SVM* może być sformułowany jako następujący problem optymalizacji z ograniczeniami:

$$\begin{aligned} &\max_{\mathbf{a}} . \ \sum_{n=1}^N a_n - \frac{1}{2}\sum_{n=1}^N\sum_{m=1}^N a_n a_m y_n y_m k(\boldsymbol{\phi}^n, \boldsymbol{\phi}^m), \\ &\text{p.o.} \quad 0 \leqslant a_n \leqslant C, \\ &\qquad \sum_{n=1}^N a_n y_n = 0, \end{aligned} \tag{5.30}$$

gdzie C jest parametrem *regularyzacji* klasyfikatora i steruje jego pojemnością. Innymi słowy, im wyższe C, tym klasyfikator posiada wyższy *wymiar Vapnika-Chervonenkisa* (ang.

VC-dimension) i może separować szerszą klasę zbiorów [135, 165, 166]. Funkcja $k(\cdot,\cdot)$ jest funkcją jądra (ang. kernel) i należy ją interpretować jako miara podobieństwa pomiędzy deskryptorami. W pracy użyto jądro gaussowskie:

$$k(\boldsymbol{\phi}^n, \boldsymbol{\phi}^m) = \exp\left(\frac{1}{2\sigma_\phi^2}\|\boldsymbol{\phi}^n - \boldsymbol{\phi}^m\|^2\right), \tag{5.31}$$

gdzie parametr σ_ϕ steruje precyzją jądra. Należy zauważyć, że sformułowany problem optymalizacji (5.30) jest przykładem *programowania kwadratowego* (ang. quadratic programming) i zalicza się do klasy problemów wypukłych, które mogą być efektywnie rozwiązywane. Dedykowaną metodą do uczenia *SVM*, rozwiązującą ten problem, jest algorytm *Sequential Minimal Optimization* (SMO) [122], który jest bardzo wydajny nawet dla dużych i wysokowymiarowych zbiorów przykładów. Niemniej jednak problem (5.30) może być rozwiązany z użyciem standardowym algorytmów do optymalizacji wypukłej, np. *Interior-Point* [21].

Wynikiem uczenia klasyfikatora jest wektor parametrów $\mathbf{a}$, który często jest wektorem rzadkim[3] (ang. sparse), tj. posiada wiele zerowych elementów. Przez SV oznaczmy zbiór indeksów n, dla których a_n są niezerowe. Odpowiadające im przykłady $\boldsymbol{\phi}^n$ nazywa się wektorami wspierającymi (ang. support vector). Siła reakcji modelu na nieobserwowany dotąd przykład $\boldsymbol{\phi}$ wyraża się następującą zależnością:

$$s(\boldsymbol{\phi}) = \sum_{n\in SV} a_n y_n k(\boldsymbol{\phi}, \boldsymbol{\phi}^n) + b, \tag{5.32}$$

gdzie b oznacza przesunięcie (ang. bias) i może być wyznaczone w następujący sposób:

$$b = \frac{1}{N_{SV}} \sum_{m\in SV} \sum_{n\in SV} a_n y_n k(\boldsymbol{\phi}^m, \boldsymbol{\phi}^n) - y_m. \tag{5.33}$$

Reguła klasyfikacyjna, tj. zasada według której nowym obserwacjom nadawana jest etykieta klasy $y \in \{-1, 1\}$, polega na ustaleniu progu s_0[4] takiego, że dla $s(\boldsymbol{\phi}) > s_0$ obserwacje klasyfikowane są do jednej klasy, a dla $s(\boldsymbol{\phi}) \leqslant s_0$ do drugiej.

Na *model wyglądu* składają się funkcje $s_k(\boldsymbol{\phi})$ o postaci (5.32) zdefiniowane odpowiednio dla każdego punktu o unikatowym indeksie k wyróżnionym na *modelu ciała*. Warto zauwa-

[3] Przede wszystkim w przypadku niskowymiarowych danych, dla których płaszczyzna separująca nie jest silnie nieliniowa.

[4] Zazwyczaj przyjmuje się, że $s_0 = 0$.

żyć, że im wyższa wartość tych funkcji, tym punkt na obrazie (opisany przez deskryptor) jest bardziej podobny do punktu ustalonego na *modelu ciała*.

Na koniec warto podać kilka uwag dotyczących *modelu wyglądu*:

1. Problem uczenia *SVM* w postaci (5.30) jest problemem dualnym do problemu uczenia klasyfikatora z miękkim marginesem (ang. soft-margin classifier). W literaturze dotyczącej *SVM* najpierw formułowany jest problem prymalny, który jest bardziej intuicyjny i ma prostą geometryczną interpretację. Następnie korzystając z metody Lagrange'a formułuje się problem dualny i zastępuje liniowy iloczyn skalarny funkcją jądra[5]. Szczegóły tych przekształceń zawarte są w książce [135].

2. Ze względu na fakt, że deskryptory są wysokowymiarowe (od kilkuset do kilku tysięcy wymiarów), to w wyniku uczenia otrzymujemy zawsze dużą liczbę *wektorów wspierających*. To wiąże się ze znacznym kosztem obliczeniowym użycia klasyfikatora (5.32), ponieważ do sklasyfikowania pojedynczego przykładu konieczne jest wyliczenie funkcji jądra (5.31) z każdym *wektorem wspierającym*. Problem ten może być złagodzony poprzez metody *automatycznego ustalania istotności* (ang. Automatic Relevance Determination) polegającej na dołożeniu dodatkowej *regularyzacji* na parametry a_n w procesie uczenie *SVM*. Prowadzi to do modelu *Relevance Vector Machine* (RVM) [156], który charakteryzuje się znacznie mniejszą liczbą *wektorów wspierających* ustalanych w procesie uczenia, a jakością zbliżoną do *SVM*.

3. Do stworzenia *modelu wiarygodności* opartego na lokalnych deskryptorach konieczne jest, aby klasyfikator zwracał siłę reakcji na bieżący deskryptor wyrażoną w postaci liczby rzeczywistej (5.32). W związku z tym *SVM* jest jednym z wielu możliwych binarnych klasyfikatorów, które można zastosować do stworzenia *modelu wyglądu*. Przykładowo można użyć *regresji logistycznej* [17], *AdaBoost* [57], *Random Forests* [23], *Graph-based Rules Inducer* (GRI)[6] [158].

[5] Zastąpienie liniowego iloczynu skalarnego poprzez nieliniową funkcję jądra nazywa się sztuczką z funkcją jądra (ang. kernel trick).

[6] Użycie klasyfikatora GRI wymagałoby dodatkowo dyskretyzacji deskryptora ϕ.

5.4.3 Funkcja wiarygodności

Przed zdefiniowaniem *modelu wiarygodności* wprowadźmy pojęcie *funkcji sigmoidalnej*:

$$\sigma(s) = \frac{1}{1 + \exp(-s)}. \tag{5.34}$$

Powyższa funkcja ma tę własność, że przekształca prostą rzeczywistą na odcinek $[0, 1]$ i wykorzystana zostania do ograniczenia siły reakcji klasyfikatora (5.32). Należy zauważyć, że w przypadku dotychczas zdefiniowanych wiarygodności (5.11) i (5.24), wartości wyliczane w pojedynczych punktach na obrazie również były ograniczone do przedziału $[0, 1]$.

Ponadto wprowadźmy analogiczny wskaźnik do (5.14), który będzie oceniał czy punkt brany do wyliczenia wiarygodności jest widoczny z danej perspektywy:

$$v^{\mathrm{I}}_{ijk} = \begin{cases} 1, & \exists_l \ \ (i,j,k) \in \mathcal{M}^{\mathrm{I}}(\mathcal{V}_l) \ \wedge \ \mathrm{D}^{\mathrm{I}}_{ij} = l \\ 0, & \text{w przeciwnym przypadku} \end{cases}. \tag{5.35}$$

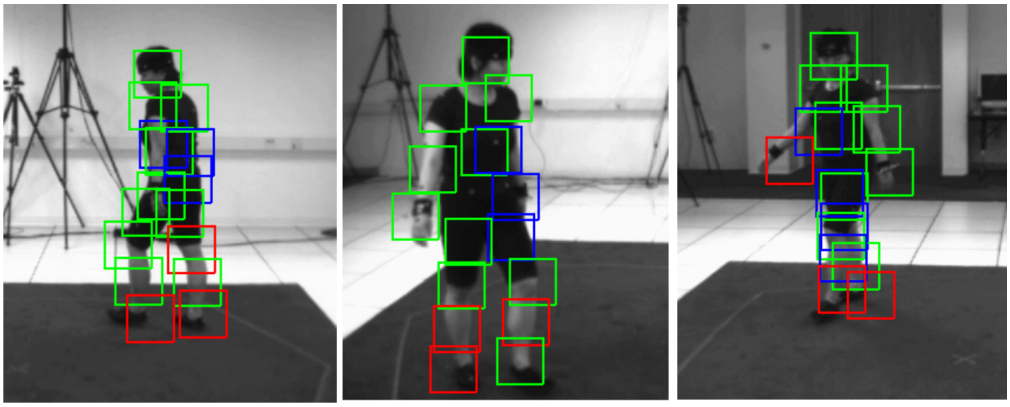
Wówczas *model wiarygodności* oparty na lokalnych deskryptorach ma następującą postać:

$$\begin{aligned} -\ln p(\mathcal{I}|\mathbf{x}) = \frac{1}{|\mathcal{I}|} \sum_{\mathrm{I} \in \mathcal{I}} \Bigg\{ \frac{1}{|\mathcal{M}^{\mathrm{I}}(\mathbf{x})|} \sum_{(i,j,k) \in \mathcal{M}^{\mathrm{I}}(\mathbf{x})} \Big[v^{\mathrm{I}}_{ijk} \Big(1 - \sigma(s_k(\boldsymbol{\phi}_{ij})) \Big) \\ + (1 - v^{\mathrm{I}}_{ijk}) \sigma_0 \Big] \Bigg\} + \mathrm{const}, \end{aligned} \tag{5.36}$$

gdzie $s_k(\boldsymbol{\phi}_{ij})$ są siłami odpowiedzi (5.32) klasyfikatora modelującego punkt o indeksie k dla deskryptora wyliczonego w punkcie (i, j) na obrazie I. Parametr σ_0 jest analogicznym współczynnikiem do e_0 w przypadku modelu (5.24) i ma na celu modelowanie wiarygodności niewidocznego punktu. Ponadto model ponownie został zdefiniowany z dokładnością do stałej normującej.

Na rysunku 5.5 zostało przedstawione wyliczenie wiarygodności dla jednego hipotetycznego *wektora stanu* i obrazów z trzech kamer. Na zielono zostały zaznaczone widoczne punkty dla których wartość $\sigma(s_k(\boldsymbol{\phi}_{ij})) > 1 - \sigma_0$, a na czerwono te, dla których ta wartość była mniejsza. Na niebiesko zaznaczono niewidoczne punkty z danej perspektywy.

Podsumowując, warto wymienić kilka uwag dotyczących *modeli wiarygodności* bazujących na wyróżnionych punktach na ciele oraz stosowania lokalnych deskryptorów do wyliczania statystyk z obrazu:



Rysunek 5.5: Model wiarygodności oparty na lokalnych deskryptorach.

1. Funkcja wiarygodności (5.36) powinna charakteryzować się tym, że nie ma nagłych skoków, jeśli zrzutowane punkty są lekko przesunięte względem ich prawidłowego położenia. Można to uzyskać poprzez odpowiednie generowanie pozytywnych przykładów do ciągów treningowych (5.29). Zamiast wybierać do niego jedynie deskryptory wyliczone w wyróżnionych punktach, należy losować (np. z rozkładu normalnego) punkty wokół i wyliczone w nich deskryptory również dodawać do ciągu uczącego. Dodatkowo można stosować rozmycie wejściowego obrazu filtrem gaussowskim, gdyż powoduje to efekt interpolacji poszczególnych pikseli, bazując na pikselach sąsiednich, a w konsekwencji zwiększa także gładkość funkcji wiarygodności.

2. W literaturze spotyka się podejście, że wyróżnione punkty są lokalizowane bezpośrednio na obrazach. Przykładowo może się to odbywać poprzez ich śledzenie [9, 40, 161], np. z użyciem metody śledzącej punkty *Wandering-Stable-Lost* [78], lub poprzez ich detekcję [128]. Wtedy wyliczenie wiarygodności odbywa się na zasadzie porównania odległości punktów wyróżnionych na *modelu ciała* i punktów wykrytych na obrazie. Należy podkreślić, że te metody są obciążone błędem wynikającym z niedoskonałości algorytmów lokalizujących punkty na obrazie.

3. Lokalne deskryptory dla wyróżnionych części ciała są powszechnie stosowane do konstruowania detektorów, wykrywających te części na obrazie. Stanowi to punkt wyjścia najskuteczniejszych obecnie metod rozwiązujących statyczny problem *estymacji*

pozy z użyciem *modeli opartych na częściach* (ang. part-based model). Bazując na informacji od detektorów, poprawiana jest wiedza aprioryczna o konfiguracji ciała i z wykorzystaniem różnych technik wnioskujących, szukane jest maksimum rozkładu a posteriori (estymator MAP) [5, 6, 7, 16, 40, 47, 95, 96, 134, 142, 143, 148, 149, 172]. Jest to przykład podejścia określanego w literaturze jako *bottom-up*, gdzie najpierw brana jest informacja z obrazu i po jej uwzględnieniu przeszukiwana jest przestrzeń stanów[7]. Podejście stosowane w pracy, określane jest jako *top-down*, gdzie najpierw przeszukuje się przestrzeń stanów (generując potencjalne konfiguracje), a następnie uwzględnia się informację o obrazie.

5.5 Łączenie modeli wiarygodności

Pojedynczy *model wiarygodności* często dostarcza jedyne część informacji o bieżącej obserwacji, która może być użyta do oceny dopasowania ustalonego *wektora stanu*. Dlatego można łączyć wiele różnych *modeli wiarygodności* w celu uwzględnienia większej ilości informacji w stosowanym schemacie *śledzenia ruchu człowieka*, co zostało wykorzystane m.in. w pracach [32, 41, 44, 97, 140].

Załóżmy, że dysponujemy L *modelami wiarygodności*, wtedy na ich podstawie możemy zbudować model, który łączy ze sobą wnoszoną przez nie informację:

$$\ln p(\mathcal{I}|\mathbf{x}) = \sum_{l=1}^{L} \lambda_l \ln p_l(\mathcal{I}|\mathbf{x}) + \text{const.} \tag{5.37}$$

Parametry $\lambda_l \geqslant 0$ oznaczają wagi, które decydują o sile, z jaką informacja z danego modelu jest uwzględniana w całościowej wiarygodności. Istotne jest, aby składowe modele (w zlogarytmowanej postaci) dawały wartości liczbowe, które są ze sobą porównywalne. Należy podkreślić, że przedstawione w tym rozdziale *modele wiarygodności* (5.11), (5.24), (5.36) zwracają wartości z przedziału $[0, 1]$.

[7] W *modelach opartych na częściach wektor stanu* złożony jest często z położenia poszczególnych części na obrazie, a nie jak w pracy – ze względnych obrotów.

Rozdział 6

Modele dynamiki

W rozdziale zostały przedstawione dwa *modele dynamiki*. Pierwszy znany z literatury, stosowany do problemu *śledzania ruchu* (3.16). Drugi autorski, stosowany do problemu *śledzenia ruchu* z uwzględnieniem niskowymiarowej rozmaitości (3.24).

6.1 Model podstawowy

W praktyce często stosuje się prosty *model dynamiki* o następującej postaci [44, 140]:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{6.1}$$

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{\varepsilon}_t|0, \mathrm{diag}(\boldsymbol{\sigma}^2)), \tag{6.2}$$

gdzie zakłada się, że kolejny stan jest przewidywany na bazie poprzedniego zmodyfikowanego poprzez dodanie szumu gaussowskiego. Wtedy warunkowy rozkład na na *wektor stanu* pojawiający się w problemie (3.16) jest postaci:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2)), \tag{6.3}$$

tj. ma rozkład normalny o średniej $\mathbf{x}_{t-1}$ i diagonalnej macierzy kowariancji $\mathrm{diag}(\boldsymbol{\sigma}^2)$.

Parametry modelu mogą być wyznaczone na podstawie ciągu treningowego *wektorów stanu* $\mathcal{D}_x = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ uzyskanego przy pomocy *systemu MOCAP*. Najbardziej naturalnym podejściem jest zastosowanie estymatora największej wiarygodności niezależnie do

każdego elementu macierzy kowariancji:

$$\sigma_i^2 = \frac{1}{T-1}\sum_{t=2}^{T}\left(x_t^i - x_{t-1}^i\right)^2, \tag{6.4}$$

gdzie x_t^i oznacza i-tą składową wektora $\mathbf{x}_t$ zdefiniowanego przy pomocy zależności (2.36) lub (2.38) i zastępuje wspólnym oznaczeniem składowe przesunięcia i poszczególnych obrotów. Natomiast σ_i^2 oznaczają składowe wektora $\boldsymbol{\sigma}^2$.

Niestety, rozkład normalny jest lekkoogonowy. Oznacza to, że odstające obserwacje będą generowane niezwykle rzadko. W konsekwencji system śledzący będzie miał tendencję do gubienia śledzonego obiektu w przypadku, gdy różnica pomiędzy klatkami będzie istotna dla pewnych stopni swobody, ponieważ nie będzie generowana dostateczna liczba cząsteczek, by właściwie pokryć odległy fragment przestrzeni stanów. Problem ten może być złagodzony poprzez zastosowanie innego estymatora dla parametrów macierzy kowariancji. W tym celu zdefiniujmy dystrybuantę empiryczną dla wariancji i-tego elementu *wektora stanu*:

$$\hat{F}_i(x) = \frac{1}{T-1}\sum_{t=2}^{T}\mathbb{1}\Big\{(x_t^i - x_{t-1}^i)^2 \leqslant x\Big\}. \tag{6.5}$$

Estymator wariancji ustalamy wtedy na kwantyl rzędu $1-\alpha$:

$$\sigma_i^2 = \kappa_i(1-\alpha), \tag{6.6}$$

który zdefiniowany jest w następujący sposób za pomocą dystrybuanty empirycznej:

$$\kappa_i(\alpha) = \min_x\Big\{x : F_i(x) \geqslant \alpha\Big\}. \tag{6.7}$$

W praktyce dla niedużego α rozkłady charakteryzują się większą wariancją niż w przypadku estymatora (6.4), jeśli występował istotny odsetek obserwacji odstających w ciągu treningowym. Powoduje to, że wygenerowane cząsteczki pokrywają wtedy przestrzeń bardziej równomiernie i zmniejszają tendencję do gubienia się systemu śledzącego.

Podsumowując należy zauważyć, że przedstawiony model zakłada, że poszczególne składowe *wektora stanu* są od siebie niezależne (diagonalna macierz kowariancji). W rzeczywistości podczas ruchu składowe *wektora stanu* charakteryzują się silnymi zależnościami, które dodatkowo mają nieliniowy charakter. Oznacza to, że próbki generowane z rozkładu (6.3) będą często odpowiadały konfiguracjom niemożliwym do wystąpienia w rzeczywistości. Przykładowo, kończyny mogą się wzajemnie przenikać, stawy wyginać poza

zakres ruchomości lub mogą pojawiać się konfiguracje niecharakterystyczne dla wykonywanego rodzaju ruchu. Pokazuje to, że model jest zbyt daleko idącym uproszczeniem i rzeczywista dynamika ma dużo bardziej skomplikowany charakter. Jednym z możliwych usprawnień jest przykładowo uwzględnienie dodatkowej wiedzy apriorycznej o zakresie ruchomości poszczególnych stawów wynikającej z anatomii człowieka. Wtedy potencjalne konfiguracje wygenerowane z *modelu dynamiki*, które nie spełniają tych ograniczeń, są odrzucane. Zazwyczaj takie podejście prowadzi do poprawy skuteczności śledzenia, jednakże w pracy nie jest ono rozważane, ponieważ wymaga uwzględnienia wiedzy, która nie wynika bezpośrednio ze zbioru treningowego $\mathcal{D}_x$. W dalszej części zostanie pokazane, że zbiór uczący zawiera dostatecznie dużo informacji, żeby zbudować model, który uwzględnia skomplikowane zależności pomiędzy stopniami swobody.

6.2 Model uwzględniający strukturę rozmaitości

Założenie, że ruch człowieka odbywa się w pobliżu niskowymiarowej *rozmaitości* ma na celu uchwycenie korelacji pomiędzy stopniami swobody w *wektorze stanu*. Dzięki temu ograniczona zostanie możliwość generowania konfiguracji, które nie mogą pojawić się w rzeczywistości. Rozważając przedstawione w dalszej części metody wprowadza się następujące założenia:

1. *Wektor stanu* może być zapisany w następującej postaci:

$$\mathbf{x} = (\mathbf{u}_0, \boldsymbol{\theta}_0, \check{\mathbf{x}})^{\mathrm{T}}, \tag{6.8}$$

   gdzie $\mathbf{u}_0$ i $\boldsymbol{\theta}_0$ oznaczają odpowiednio położenie i rotację człowieka w globalnym układzie współrzędnych, a $\check{\mathbf{x}}$ jest zredukowanym *wektorem stanu*, reprezentującym wewnętrzną konfigurację człowieka niezależną od położenia na scenie.

2. Wektor $\check{\mathbf{x}}$ jest w postaci (2.37), tj. obroty reprezentowane są poprzez zredukowane *kwaterniony* zdefiniowane przy pomocy zależności (2.6). Wtedy możliwe jest porównanie dwóch zredukowanych *wektorów stanu* za pomocą metryki euklidesowej, której użycie będzie uzasadnione w dalszej części rozdziału.

Struktura niskowymiarowej rozmaitości jest związana jedynie z wewnętrzną konfiguracją człowieka, tj. w jej pobliżu rozkładają się zredukowane *wektory stanu*. Ma to sens,

ponieważ stopnie swobody w wewnętrznej konfiguracji charakteryzują się silnymi nieliniowymi zależnościami podczas ustalonego rodzaju ruchu (np. chodzenia, biegania) i dodatkowo nie są związane z globalnym położeniem i rotacją. W związku z tym *model dynamiki* w naturalny sposób ulega następującej dekompozycji:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = p(\mathbf{u}_{0,t}|\mathbf{u}_{0,t-1})p(\boldsymbol{\theta}_{0,t}|\boldsymbol{\theta}_{0,t-1})p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}), \tag{6.9}$$

gdzie przyjmujemy, że dwa pierwsze modele są postaci analogicznej do (6.3), czyli:

$$p(\mathbf{u}_{0,t}|\mathbf{u}_{0,t-1}) = \mathcal{N}(\mathbf{u}_{0,t}|\mathbf{u}_{0,t-1}, \mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{u}}^2)), \tag{6.10}$$

$$p(\boldsymbol{\theta}_{0,t}|\boldsymbol{\theta}_{0,t-1}) = \mathcal{N}(\boldsymbol{\theta}_{0,t}|\boldsymbol{\theta}_{0,t-1}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)). \tag{6.11}$$

Poszczególne wariancje $\boldsymbol{\sigma}_{\mathbf{u}}^2$ i $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2$ mogą być wyznaczone z użyciem estymatora (6.4) lub (6.6).

W związku z faktem, że jedynie wewnętrzna konfiguracja rozkłada się w pobliżu niskowymiarowej rozmaitości, dlatego dekompozycja (3.26) może być zastosowana jedynie do modelu $p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1})$, ponieważ globalne położenie i rotacja nie zależą od zmiennej $\mathbf{z}$, tj.

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{x}_{t-1})\mathrm{d}\mathbf{z}_t \\ &= p(\mathbf{u}_{0,t}|\mathbf{u}_{0,t-1})p(\boldsymbol{\theta}_{0,t}|\boldsymbol{\theta}_{0,t-1}) \int p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t|\check{\mathbf{x}}_{t-1})\mathrm{d}\mathbf{z}_t. \end{aligned} \tag{6.12}$$

W celu określenia powyższego *modelu dynamiki* konieczne jest wykonanie następujących kroków:

1. Odtworzenie struktury niskowymiarowej *rozmaitości* na podstawie zbioru treningowego $\mathcal{D}_x$ z danymi z *systemu MOCAP*.

2. Zaproponowanie modelu określającego dynamikę po *rozmaitości* $p(\mathbf{z}_t|\check{\mathbf{x}}_{t-1})$.

3. Zaproponowanie modelu określającego rozkład bieżącego stanu w oparciu o stan poprzedni i bieżącą współrzędną na *rozmaitości* $p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathbf{z}_t)$.

Śledzenie ruchu może być wtedy przeprowadzone z użyciem algorytmu 3, gdzie składowe cząsteczek $\mathbf{x}_t^{(n)}$ związane z globalnym położeniem i rotacją, mogą być generowane poprzez próbkowanie bezpośrednio z rozkładów (6.10) i (6.11).

6.2.1 Odtworzenie struktury rozmaitości

W celu odtworzenia struktury niskowymiarowej *rozmaitości* potrzebujemy znaleźć reprezentację $\mathbf{z}_t$ dla każdego zredukowanego *wektora stanu* ze zbioru treningowego $\mathcal{D}_x$. Formalnie niech $\mathbf{X}$ oznacza macierz, której kolejne wiersze zawierają zredukowane *wektory stanu* ze zbioru uczącego. Wtedy odtworzenie niskowymiarowej reprezentacji polega na wyznaczeniu macierzy $\mathbf{Z}$, w której wiersze zawierają współrzędne punktów w układzie związanym z *rozmaitością*. Do znalezienia tej macierzy zostanie wykorzystany model *Gaussian Process Latent Variable Model* (GPLVM) [90].

Model ten zakłada następującą relację pomiędzy nisko i wysokowymiarowymi reprezentacjami:

$$\check{\mathbf{x}} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}, \tag{6.13}$$

$$f_i \sim \mathcal{GP}(f|0, k_z(\mathbf{z}, \mathbf{z}')) \tag{6.14}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|0, \sigma_z^2 \mathbf{I}_{D\times D}), \tag{6.15}$$

gdzie D oznacza wymiar zredukowanego *wektora stanu*. Zmienna losowa $\boldsymbol{\varepsilon}$ jest szumem gaussowskim o D wymiarach i jednakowej wariancji σ_z^2 dla wszystkich współrzędnych. Zapis $\mathbf{f}(\mathbf{z})$ oznacza D-wymiarowy wektor funkcji o składowych $f_i(\mathbf{z})$, z których każda z nich jest niezależną realizacją *procesu Gaussa* (ang. Gaussian process) o średniej zero i funkcji kowariancji $k_z(\mathbf{z}, \mathbf{z}')$. Potraktowanie odwzorowania między zmiennymi jako realizacji *procesu Gaussa* pozwala na modelowanie silnie nieliniowych zależności poprzez odpowiedni dobór funkcji kowariancji i jest obecnie podstawowym narzędziem w *uczeniu maszynowym* do rozwiązywania problemu regresji. Więcej o technikach wykorzystujących *procesy Gaussa* można znaleźć w książce [127]. W pracy została przyjęta następująca funkcja kowariancji:

$$k_z(\mathbf{z}, \mathbf{z}') = \beta \exp\left(-\frac{\gamma_z}{2}\|\mathbf{z} - \mathbf{z}'\|^2\right) + \beta_0, \tag{6.16}$$

gdzie β, β_0 i γ_z są pewnymi ustalonymi parametrami. Parametr γ_z steruje precyzją jądra gaussowskiego, a w konsekwencji odpowiada za gładkość funkcji f_i, tj. im niższa jego wartość, tym funkcja jest bardziej gładka.

W celu wyznaczenia macierzy $\mathbf{Z}$ zdefiniujemy funkcję wiarygodności $p(\mathbf{X}|\mathbf{Z})$, a następnie poprzez jej maksymalizację otrzymamy estymator największej wiarygodności dla $\mathbf{Z}$.

Oznaczmy przez $\mathbf{f}_i = (f_i(\mathbf{z}_1), \ldots, f_i(\mathbf{z}_T))^{\mathrm{T}}$ wektor zawierający wartości funkcji f_i odpowiednio dla zmiennych $\mathbf{z}_t$ oznaczających niskowymiarowe współrzędne dla elementów ze zbioru treningowego. Ponadto zdefiniujmy następującą macierz:

$$\mathbf{F} = [\mathbf{f}_1 \ \ldots \ \mathbf{f}_D], \tag{6.17}$$

która zawiera wartości dla funkcji odpowiadających kolejnym wymiarom wektora $\check{\mathbf{x}}$. Z definicji *procesu Gaussa* wiemy, że jeśli funkcja f_i jest realizacją $\mathcal{GP}(f|0, k_z(\mathbf{z}_n, \mathbf{z}_m))$, to w skończonej liczbie punktów $\mathbf{f}_i$ ma następujący rozkład:

$$p(\mathbf{f}_i|\mathbf{Z}) = \mathcal{N}(\mathbf{f}_i|0, \mathbf{K}), \tag{6.18}$$

gdzie elementy macierzy kowariancji $\mathbf{K} = [k_{nm}]$ określone są na podstawie funkcji (6.16):

$$k_{nm} = k_z(\mathbf{z}_n, \mathbf{z}_m), \tag{6.19}$$

dla wszystkich przykładów z macierzy $\mathbf{Z}$, tj. $n, m = 1, \ldots, T$. Funkcje f_i są niezależnymi realizacjami *procesu Gaussa*, a zatem rozkład dla całej macierzy (6.17) ma postać:

$$p(\mathbf{F}|\mathbf{Z}) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{f}_i|0, \mathbf{K}). \tag{6.20}$$

Dalej zakładamy, że dla kolejnych przykładów z macierzy $\mathbf{X}$ realizacje szumu (6.15) są niezależne, co prowadzi do następującego rozkładu:

$$\begin{aligned} p(\mathbf{X}|\mathbf{F}) &= \prod_{t=1}^{T} \mathcal{N}(\check{\mathbf{x}}_t|\mathbf{f}(\mathbf{z}_t), \sigma_z^2 \mathbf{I}_{D\times D}) \\ &= \prod_{t=1}^{T}\prod_{i=1}^{D} \mathcal{N}(\check{x}_t^i|f_i(\mathbf{z}_t), \sigma_z^2), \end{aligned} \tag{6.21}$$

gdzie druga równość wynika z faktu, że macierz kowariancji jest diagonalna, tj. kolejne wymiary są od siebie niezależne. Należy podkreślić, że wartości $f_i(\mathbf{z}_t)$ są elementami macierzy (6.17). W oparciu o zależności (6.20) i (6.21) możemy zdefiniować funkcję wiarygodności:

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z}) &= \int p(\mathbf{X}|\mathbf{F}) p(\mathbf{F}|\mathbf{Z}) \mathrm{d}\mathbf{F} \\ &= \int \prod_{t=1}^{T}\prod_{i=1}^{D} \mathcal{N}(\check{x}_t^i|f_i(\mathbf{z}_t), \sigma_z^2) \mathcal{N}(\mathbf{f}_i|0, \mathbf{K}) \mathrm{d}\mathbf{F} \\ &= \prod_{i=1}^{D} \int \mathcal{N}(\mathbf{X}_{:,i}|\mathbf{f}_i, \sigma_z^2 \mathbf{I}_{T\times T}) \mathcal{N}(\mathbf{f}_i|0, \mathbf{K}) \mathrm{d}\mathbf{f}_i \\ &= \prod_{i=1}^{D} \mathcal{N}(\mathbf{X}_{:,i}|0, \mathbf{K} + \sigma_z^2 \mathbf{I}_{T\times T}), \end{aligned} \tag{6.22}$$

gdzie zapis $\mathbf{X}_{:,i}$ oznacza i-tą kolumnę macierzy $\mathbf{X}$. Powyższa kombinacja dwóch rozkładów normalnych jest przykładem *liniowego modelu gaussowskiego* (ang. linear Gaussian model) i całkowanie po $\mathbf{F}$ może być wyznaczone analitycznie, wykorzystując ogólne własności dla tych modeli [132].

W celu maksymalizowania funkcji wiarygodności (6.22) skorzystamy z jej zlogarytmowanej postaci:

$$\begin{aligned}\ln p(\mathbf{X}|\mathbf{Z}) &= \sum_{i=1}^{D} \ln \mathcal{N}(\mathbf{X}_{:,i}|0, \mathbf{K} + \sigma_z^2 \mathbf{I}_{T\times T}) \\ &= \sum_{i=1}^{D} \left\{ -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\overline{\mathbf{K}}| - \frac{1}{2}\mathbf{X}_{:,i}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\mathbf{X}_{:,i} \right\} \\ &= -\frac{DT}{2}\ln(2\pi) - \frac{D}{2}\ln|\overline{\mathbf{K}}| - \frac{1}{2}\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\mathbf{X}), \end{aligned} \tag{6.23}$$

gdzie $|\cdot|$ i $\mathrm{tr}(\cdot)$ oznaczają odpowiednio wyznacznik i ślad macierzy oraz $\overline{\mathbf{K}} = \mathbf{K} + \sigma_z^2\mathbf{I}_{T\times T}$. Powyższą funkcję maksymalizujemy ze względu na macierz $\mathbf{Z}$ oraz parametry funkcji kowariancji β, β_0, γ_z i wariancję szumu σ_z^2. Należy zauważyć, że dla powyższej postaci istnieje niejednoznaczność rozwiązania ze względu na postać funkcji kowariancji (6.16), gdzie optymalizujemy jednocześnie zmienne $\mathbf{z}_t$ i parametr γ_z, które mogą być dowolnie przeskalowane, nie zmieniając wartości funkcji celu. Aby wyeliminować tę niejednoznaczność od funkcji celu odejmuje się *regularyzator* o następującej postaci [1]:

$$\frac{1}{2}\|\mathbf{Z}\|_F^2, \tag{6.24}$$

gdzie $\|\cdot\|_F$ oznacza normę Frobeniusa dla macierzy. Ostatecznie otrzymujemy następującą postać funkcji celu:

$$L(\mathbf{Z}) = \ln p(\mathbf{X}|\mathbf{Z}) - \frac{1}{2}\|\mathbf{Z}\|_F^2. \tag{6.25}$$

Powyższa funkcja może być optymalizowana z użyciem standardowych technik do optymalizacji numerycznej funkcji ciągłych, jak *metoda gradientów sprzężonych* (ang. conjugate gradient method) lub *algorytm Broydena-Fletchera-Goldfarba-Shanno* (BFGS) [115]. Należy zauważyć, że ze względu na występowanie funkcji kowariancji, rozważana funkcja nie jest

[1] Wprowadzenie *regularyzatora* o postaci (6.24) jest równoważne z nałożeniem rozkładu a priori na macierz $\mathbf{Z}$ o postaci $p(\mathbf{Z}) = \prod_{t=1}^{T}\mathcal{N}(\mathbf{z}_t|0, \mathbf{I}_{d\times d})$.

wklęsła, a w konsekwencji posiada liczne maksima lokalne. Dlatego konieczne jest, aby algorytmy numerycznej optymalizacji były właściwie zainicjalizowane. Początkowe wartości macierzy $\mathbf{Z}$ ustala się przykładowo przy użyciu metody *analizy głównych składowych* (ang. principal component analysis)[1].

Algorytmy optymalizacji potrzebują informacji o gradiencie funkcji (6.25). Innymi słowy, konieczne jest wyznaczenie pochodnych cząstkowych względem każdego wymiaru każdej zmiennej $\mathbf{z}_t$ oraz względem pozostałych parametrów. Do ich efektywnego wyliczenia wykorzystane zostaną własności rachunku różniczkowego na wektorach i macierzach [118]. Wyznaczmy najpierw gradient funkcji (6.23) względem macierzy $\overline{\mathbf{K}}$:

$$\frac{\partial \ln p(\mathbf{X}|\mathbf{Z})}{\partial \overline{\mathbf{K}}} = -\frac{D}{2}\overline{\mathbf{K}}^{-1} + \frac{1}{2}\overline{\mathbf{K}}^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}, \tag{6.26}$$

gdzie powyższy zapis oznacza macierz o wymiarach $T \times T$ z elementami $\frac{\partial \ln p(\mathbf{X}|\mathbf{Z})}{\partial \overline{k}_{nm}}$. Warto dodać, że powyższy gradient jest niezależny od postaci macierzy $\overline{\mathbf{K}}$, a w konsekwencji niezależny od wyboru funkcji kowariancji. Następnie możemy wyliczyć pochodne elementu macierzy względem składowych z_t^i:

$$\frac{\partial \overline{k}_z(\mathbf{z}_n, \mathbf{z}_m)}{\partial z_t^i} = \begin{cases} -\gamma_z(k_{nm} - \beta_0)(z_n^i - z_m^i), & t = n \\ \gamma_z(k_{nm} - \beta_0)(z_n^i - z_m^i), & t = m \\ 0, & t \neq n, m \end{cases}, \tag{6.27}$$

gdzie funkcja $\overline{k}_z(\mathbf{z}_n, \mathbf{z}_m)$ została zdefiniowana następująco:

$$\overline{k}_z(\mathbf{z}_n, \mathbf{z}_m) = k_z(\mathbf{z}_n, \mathbf{z}_m) + \sigma_z^2 \delta_{nm} \tag{6.28}$$

i wyraża, w jaki sposób wyliczane są elementy macierz $\overline{\mathbf{K}}$, gdzie δ_{nm} oznacza deltę Kroneckera. Na koniec wyliczamy pochodną z *regularyzatora* (6.24) względem składowych z_t^i:

$$\frac{\partial \|\mathbf{Z}\|_F^2}{\partial z_t^i} = 2z_t^i. \tag{6.29}$$

Zbierając razem (6.26), (6.27) i (6.29) oraz korzystając z reguły łańcuchowej dla pochodnych cząstkowych, otrzymujemy wyrażenie na pochodną funkcji celu (6.25) względem z_t^i:

$$\frac{\partial L}{\partial z_t^i} = \mathrm{tr}\left(\left(\frac{\partial \ln p(\mathbf{X}|\mathbf{Z})}{\partial \overline{\mathbf{K}}}\right)^{\mathrm{T}} \frac{\partial \overline{\mathbf{K}}}{\partial z_t^i}\right) - \frac{1}{2}\frac{\partial \|\mathbf{Z}\|_F^2}{\partial z_t^i}, \tag{6.30}$$

gdzie elementy macierzy $\frac{\partial \overline{\mathbf{K}}}{\partial z_t^i}$ są postaci (6.27). W analogiczny sposób możemy wyznaczyć pochodne względem parametrów β, β_0, γ_z i σ_z^2 zastępując (6.27) pochodnymi względem odpowiednich parametrów.

Należy zauważyć, że funkcja kowariancji (6.16) przyjmuje tym wyższe wartości, im dwa punkty $\mathbf{z}_n$ i $\mathbf{z}_m$ są bliżej siebie położone. Innymi słowy, tj. są bardziej do siebie podobne. Ponieważ funkcja ta definiuje macierz kowariancji dla rozkładu (6.22), oznacza to, że dla sąsiednich punktów $\mathbf{z}_n$ i $\mathbf{z}_m$ odpowiadające im zredukowane *wektory stanu* $\check{\mathbf{x}}_n$ i $\check{\mathbf{x}}_m$ są silnie skorelowane, a konsekwencji również do siebie podobne na wszystkich składowych jednocześnie. Sytuacja ta nie zachodzi w drugą stronę, tj. jeśli dwa punkty $\check{\mathbf{x}}_n$ i $\check{\mathbf{x}}_m$ położone są blisko siebie, to odpowiadające im współrzędne niskowymiarowe $\mathbf{z}_n$ i $\mathbf{z}_m$ mogą być od siebie odległe. Intuicyjnie zachodzi tak wtedy, jeśli *rozmaitość* jest silnie pozwijana w przestrzeni wysokowymiarowej. W problemie rozważanym w pracy sytuacja ta jest niekorzystna, ponieważ wtedy postać rozkładu $p(\mathbf{z}_t|\check{\mathbf{x}}_{t-1})$ będzie wielomodalna i trudna do wyznaczenia. Efekt ten można złagodzić, używając tzw. *ograniczeń wstecznych* (ang. back constraints). Prowadzi to do modelu *Back-Constrained Gaussian Process Latent Variable Model* (BC-GPLVM) [92].

Idea tego modelu polega na tym, aby zmienną $\mathbf{z}$ zdefiniować jako gładkie odwzorowanie zmiennej $\check{\mathbf{x}}$:

$$\mathbf{z} = \mathbf{g}(\check{\mathbf{x}}). \tag{6.31}$$

Przez pojęcie gładkości rozumiemy fakt, że odwzorowanie nie posiada gwałtownych zmian w otoczeniu dowolnego $\check{\mathbf{x}}$, co w konsekwencji prowadzi do własności, że dla podobnych $\check{\mathbf{x}}_n$, $\check{\mathbf{x}}_m$ otrzymamy podobne $\mathbf{z}_n$, $\mathbf{z}_m$. Przykładowo odwzorowanie (6.31) można zdefiniować w postaci liniowego modelu z cechami zadanymi przez funkcję jądra:

$$g_i(\check{\mathbf{x}}) = \sum_{t=1}^{T} c_{ti} k_x(\check{\mathbf{x}}, \check{\mathbf{x}}_t) + b_i, \tag{6.32}$$

gdzie g_i oznacza składową odpowiadającą i-tej współrzędnej wektora $\mathbf{z}$, dla $i = 1, \ldots, d$. Funkcje g_i są sparametryzowane przez współczynniki c_{ti} i przesunięcie b_i. Do określenia funkcji jądra została wykorzystana analogiczna postać do (6.16):

$$k_x(\check{\mathbf{x}}, \check{\mathbf{x}}') = \exp\left(-\frac{\gamma_x}{2}\|\check{\mathbf{x}} - \check{\mathbf{x}}'\|^2\right). \tag{6.33}$$

Parametr γ_x określa precyzję powyższej funkcji i steruje gładkością odwzorowania (6.31), tj. im niższa jego wartość, tym odwzorowania jest bardziej gładkie. W tym miejscu należy

podkreślić, że aby funkcja jądra prawidłowo określała podobieństwo pomiędzy zredukowanymi *wektorami stanu*, konieczne jest by wektory mogły być porównywane przy pomocy metryki euklidesowej. Stąd wynika przyjęte w rozdziale założenie o składowych wektora $\check{\mathbf{x}}$ w postaci zredukowanych *kwaternionów*.

Odwzorowanie (6.31) możemy podstawić do funkcji celu (6.25), tj. za każdą składową niskowymiarowych wektorów podstawić $z_n^i = g_i(\check{\mathbf{x}}_n)$, a następnie optymalizować tę funkcję względem zmiennych c_{ti} i b_i, zamiast względem współrzędnych z_n^i. Można to zrobić z użyciem tych samych metod optymalizacji, co w przypadku (6.25). Do wyznaczenia składowych gradientu dla parametrów c_{ti} należy zastosować regułę łańcuchową do wyrażenia (6.30):

$$\frac{\partial L}{\partial c_{ti}} = \sum_{n=1}^{T} \frac{\partial L}{\partial z_n^i} \frac{\partial z_n^i}{\partial c_{ti}}, \tag{6.34}$$

gdzie wartości $\frac{\partial z_n^i}{\partial c_{ti}}$ wynikają natychmiast z postaci funkcji (6.32):

$$\frac{\partial z_n^i}{\partial c_{ti}} = k_x(\check{\mathbf{x}}_n, \check{\mathbf{x}}_t). \tag{6.35}$$

W analogiczny sposób możemy wyznaczyć składowe gradientu dla parametrów b_i.

Użycie *ograniczeń wstecznych* powoduje, że otrzymane niskowymiarowe reprezentacje $\mathbf{z}_t$ odpowiadające punktom $\check{\mathbf{x}}_t$ ze zbioru treningowego położone są blisko siebie, jeśli wysokowymiarowe wektory są do siebie podobne.

Podsumowując, warto wymienić kilka uwag dotyczących sposobu odtwarzania struktury *rozmaitości* z użyciem modelu *GPLVM*:

1. Rozkład (6.22) w naturalny sposób może być rozszerzony o nową parę $(\check{\mathbf{x}}, \mathbf{z})$, gdzie składowe rozkłady normalne będą sparametryzowane macierzą kowariancji o następującej postaci blokowej:

$$\begin{bmatrix} \overline{\mathbf{K}} & \overline{\mathbf{k}} \\ \overline{\mathbf{k}}^{\mathrm{T}} & \overline{k}_z(\mathbf{z}, \mathbf{z}) \end{bmatrix}, \tag{6.36}$$

   w której elementy wektora $\overline{\mathbf{k}}$ zdefiniowane są przy pomocy funkcji kowariancji (6.28), tj. $\overline{k}_t = \overline{k}_z(\mathbf{z}, \mathbf{z}_t)$. Następnie korzystając z własności wielowymiarowych rozkładów normalnych [17], możemy przenieść macierz $\mathbf{X}$ do warunku i wyznaczyć *rozkład predykcyjny* (ang. predictive distribution):

$$p(\check{\mathbf{x}}|\mathbf{z}, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\check{\mathbf{x}}|\boldsymbol{\mu}_p, \sigma_p^2 \mathbf{I}_{D\times D}), \tag{6.37}$$

gdzie wartości parametrów wynikają bezpośrednio z ogólnej postaci warunkowych rozkładów normalnych:

$$\boldsymbol{\mu}_p = \mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \tag{6.38}$$

$$\sigma_p^2 = \overline{k}_z(\mathbf{z},\mathbf{z}) - \overline{\mathbf{k}}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}. \tag{6.39}$$

Zauważmy, że (6.37) definiuje rozkład na $\check{\mathbf{x}}$ przy założeniu znajomości położenia na *rozmaitości* $\mathbf{z}$ oraz struktury *rozmaitości* określonej poprzez odpowiadające sobie punkty ze zbioru treningowego wyrażone w wysokowymiarowej reprezentacji $\mathbf{X}$ i niskowymiarowej $\mathbf{Z}$.

2. Optymalizacja funkcji celu (6.25) wymaga w każdej iteracji odwrócenia macierzy $\overline{\mathbf{K}}$ w celu wyliczenia gradientu (6.26). Wiąże się to z kosztem obliczeniowym rzędu $O(T^3)$, który ogranicza zasięg stosowania tej procedury do zbiorów treningowych zawierających co najwyżej kilku tysiącach przykładów. Problem ten może być złagodzony poprzez zastosowanie technik rozrzedzających (ang. sparsification techniques) dla *procesów Gaussa*, przykładowo metody *Informative Vector Machine* (IVM) [93]. Techniki te pozwalają na wyliczanie kolejnych kroków optymalizacji na podstawie niewielkich podzbiorów zbioru uczącego.

3. W ostatnich latach powstało wiele modeli rozszerzających ideę *GPLVM*, m. in. hierarchiczny *GPLVM* [91], model *GPLVM* z ograniczeniami na topologię rozmaitości [163], model *Shared GPLVM* pozwalający na integrację danych o różnym charakterze [48], model uwzględniający dynamikę po rozmaitości – *Gaussian Process Dynamical Model* (GPDM) [169], bayesowski *GPLVM* pozwalający na automatyczny wybór wymiaru *rozmaitości* [157]. Podane rozwinięcia wskazują potencjalne kierunki dalszych prac na problemem *śledzenia ruchu człowieka*.

4. Użyta metoda wyznacza macierz $\mathbf{Z}$ poprzez maksymalizację funkcji wiarygodności $p(\mathbf{X}|\mathbf{Z})$. Należy podkreślić, że istnieją techniki do odtwarzania struktury nieliniowej *rozmaitości* korzystające z innych kryteriów. Największą grupę stanowią *metody spektralne* (ang. spectral methods) wykorzystujące dekompozycję macierzy podobieństwa na macierze wektorów i wartości własnych, zalicza się do nich m. in.: *Isomap*

[154], *Laplacian Eigenmaps* [11], *Locally Linear Embedding* [133], *Maximum Variance Unfolding* [170]. Inne techniki dopasowują parametryczne modele przybliżające zależność pomiędzy nisko i wysokowymiarowymi reprezentacjami poprzez maksymalizację funkcji wiarygodności względem parametrów, przykładowo *Generative Topographic Mapping* [18]. Jeszcze innym podejściem jest maksymalizacja informacji wzajemnej (ang. mutual information) pomiędzy nisko i wysokowymiarową reprezentacją, bazuje na tym technika *Kernel Information Embedding* [107].

5. Na koniec należy podkreślić, że model *GPLVM* i jego modyfikacje były stosowane jako fragmenty systemów śledzących ruch człowieka, m. in. w pracach [33, 48, 63, 66, 71, 155, 161, 164, 169].

6.2.2 Dynamika na rozmaitości

Idea modelu $p(\mathbf{z}_t|\check{\mathbf{x}}_{t-1})$ polega na tym, aby przewidywał on położenie w układzie współrzędnych związanym z *rozmaitością*, bazując na poprzednim położeniu w przestrzeni stanów. Zatem do jego konstrukcji potrzeba odwzorowania, które pozwoli na przejście pomiędzy reprezentacjami wysokowymiarową i niskowymiarową. W tym celu można wykorzystać *ograniczenia wsteczne* zdefiniowane zależnością (6.31), co prowadzi do następującego *modelu dynamiki* w niskowymiarowym układzie:

$$\mathbf{z}_t = \mathbf{g}(\check{\mathbf{x}}_{t-1}) + \boldsymbol{\varepsilon}_t^{x\to z}, \tag{6.40}$$

$$\boldsymbol{\varepsilon}_t^{x\to z} \sim \mathcal{N}(\boldsymbol{\varepsilon}_t^{x\to z}|0, \mathrm{diag}(\boldsymbol{\sigma}_{x\to z}^2)), \tag{6.41}$$

W powyższym modelu poprzednie położenie $\mathbf{z}_{t-1}$, odtworzone w oparciu o odwzorowanie $\mathbf{g}$, jest zaburzane addytywnym szumem gaussowskim o niezależnych składowych. Stąd rozkład zmiennej $\mathbf{z}_t$ jest postaci:

$$p(\mathbf{z}_t|\check{\mathbf{x}}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{g}(\check{\mathbf{x}}_{t-1}), \mathrm{diag}(\boldsymbol{\sigma}_{x\to z}^2)). \tag{6.42}$$

Elementy wektora $\boldsymbol{\sigma}_{x\to z}^2$ oznaczają wariancje wzdłuż poszczególnych składowych $\mathbf{z}$ i mogą być wyznaczone w oparciu o estymatory analogiczne do (6.4) lub (6.6), które zamiast elementów ciągu treningowego $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, wykorzystują niskowymiarowe współrzędne $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$, odpowiadające kolejnym wierszom macierzy $\mathbf{Z}$ wyznaczonej w procesie minimalizacji (6.25).

Należy zauważyć, że do sformułowania modelu (6.40) zostało wykorzystane odwzorowanie g, które jest konsekwencją zastosowanego do *redukcji wymiarów* modelu *BC-GPLVM* i uzasadnia jego użycie w pracy. Większość metod do *redukcji wymiarów* nie wyznacza postaci odwzorowania g, które może być użyte do rzutowania nowych obserwacji na *rozmaitość*. Niemniej odwzorowanie to może być przybliżone w oparciu o *regresję jądrową* (ang. kernel regression) [2, 60], technikę *Out-Of-Sample* [13] lub dowolny inny model nauczony na podstawie wzajemnej relacji między $\mathbf{X}$ i $\mathbf{Z}$.

6.2.3 Dynamika w przestrzeni stanów z uwzględnieniem rozmaitości

Model $p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathbf{z}_t)$ ma na celu ocenić prawdopodobieństwo stanu $\check{\mathbf{x}}_t$, bazując na stanie poprzednim oraz na bieżącym położeniu w układzie związanym z niskowymiarową *rozmaitością*. Rozsądnym założeniem jest przyjęcie, że model faktoryzuje się na dwa komponenty, z których jeden zależy od $\check{\mathbf{x}}_{t-1}$, a drugi od $\mathbf{z}_t$. Wynika to z faktu, że zmienne te należą do innych przestrzeni, a przez to ich wartości nie mogą być porównywane. Wtedy model przyjmuje postać:

$$\tilde{p}(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathbf{z}_t) = p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1})p(\check{\mathbf{x}}_t|\mathbf{z}_t). \tag{6.43}$$

Należy zauważyć, że powyższy rozkład jest nieunormowany, gdyż iloczyn dwóch rozkładów prawdopodobieństwa nie całkuje się do jedności.

Pierwszy z komponentów został przyjęty w analogicznej postaci do (6.3):

$$p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}) = \mathcal{N}(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x})), \tag{6.44}$$

gdzie zakładamy, że kolejny stan ma rozkład normalny o średniej w bieżącym stanie i diagonalnej macierzy kowariancji. Inaczej mówiąc, każdy stopień swobody może ulec drobnemu, niezależnemu zaburzeniu w stosunku do poprzedniej chwili. Poszczególne składowe wektora $\boldsymbol{\sigma}^2_{x\to x}$ mogą być wyznaczone w oparciu o ciąg treningowy $\mathcal{D}$ z użyciem estymatorów (6.4) lub (6.6).

Do zdefiniowania drugiego z komponentów wykorzystamy średnią z *rozkładu predykcyjnego* (6.37). Wtedy można zaproponować następujące odwzorowanie:

$$\check{\mathbf{x}}_t = \mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}} + \boldsymbol{\varepsilon}^{z\to x}_t, \tag{6.45}$$

$$\boldsymbol{\varepsilon}^{z\to x}_t \sim \mathcal{N}(\boldsymbol{\varepsilon}^{z\to x}_t|0, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})). \tag{6.46}$$

Elementy $\overline{k}_n$ wektora $\overline{\mathbf{k}}$ wyliczone są według następującej zależności:

$$\overline{k}_n = \overline{k}(\mathbf{z}_t, \mathbf{z}_n), \tag{6.47}$$

gdzie $\mathbf{z}_n$ stanowią wiersze macierzy $\mathbf{Z}$ wyznaczonej w procesie optymalizacji (6.25). Funkcja kowariancji $\overline{k}(\cdot,\cdot)$ określona jest przez równanie (6.28). Wynika stąd, że rozkład wektora $\check{\mathbf{x}}_t$ jest postaci:

$$p(\check{\mathbf{x}}_t|\mathbf{z}_t) = \mathcal{N}(\check{\mathbf{x}}_t|\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})). \tag{6.48}$$

Parametry $\boldsymbol{\sigma}^2_{z\to x}$ oznaczają wariancję dla poszczególnych stopni swobody wektora $\check{\mathbf{x}}_t$ i określają na ile rzeczywista konfiguracja może się odchylić od rekonstrukcji wykonanej na podstawie niskowymiarowej reprezentacji $\mathbf{z}_t$. Do ich wyznaczenia należy zastosować osobny ciąg walidacyjny, zawierający dane z *systemu MOCAP*. Wynika to z faktu, że *rozmaitość* jest wyznaczona na podstawie ciągu treningowego $\mathcal{D}_x$, a zatem rekonstrukcja w tych punktach może być precyzyjnie ustalona. W konsekwencji oszacowanie wariancji na postawie ciągu treningowego mogłoby dawać zaniżone wartości parametrów. Dla każdego przykładu z ciągu walidacyjnego wyznaczana jest niskowymiarowa reprezentacja z użyciem odwzorowania (6.31), a następnie na podstawie uzyskanej wartości rekonstruowany jest wysokowymiarowy wektor przy użyciu wyrażenia $\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}$. Poszczególne wariancje wyznaczane są przy użyciu analogicznych estymatorów do (6.4) lub (6.6), gdzie oszacowanie odbywa się na podstawie różnicy na poszczególnych stopniach swobody pomiędzy rzeczywistym wektorem z ciągu walidacyjnego i jego rekonstrukcją.

Do zastosowania algorytmu 3 konieczne jest zdefiniowanie pomocniczego rozkładu $q(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1})$, z którego generowane są hipotetyczne konfiguracje. Przyjmijmy, że:

$$q(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}) = p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}), \tag{6.49}$$

gdzie $p(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1})$ określone jest przez zależność (6.44). Wtedy współczynniki wagowe (4.27) upraszczają się do następującej postaci:

$$\begin{aligned}\tilde{\omega}(\check{\mathbf{x}}_t, \check{\mathbf{x}}_{t-1}, \mathbf{z}_t) &= \frac{\tilde{p}(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1}, \mathbf{z}_t)}{q(\check{\mathbf{x}}_t|\check{\mathbf{x}}_{t-1})} \\ &= p(\check{\mathbf{x}}_t|\mathbf{z}_t).\end{aligned} \tag{6.50}$$

Do przedstawionych rozważań należy dołączyć kilka uwag dotyczących *modelu dynamiki* uwzględniającego strukturę niskowymiarowej *rozmaitości*:

1. W przypadku stosowania innych metod *redukcji wymiarów* niż *GPLVM* może nie istnieć naturalny sposób wyznaczenia odwzorowania pomiędzy zmiennymi nisko i wysokowymiarowymi (6.45). Wtedy można je przybliżać przykładowo z użyciem *regresji jądrowej* [2, 60].

2. Kodowanie informacji o nieliniowej *rozmaitości* przy pomocy wektora o składowych rzeczywistych $\mathbf{z}$, oznaczającego położenie punktu w niskowymiarowym układzie, jest tylko jednym z możliwych podejść do tego problemu. Inne techniki wykorzystują metody *klasteryzacji*, dzieląc *rozmaitość* na drobne fragmenty, a następnie kodując wektor poprzez przynależność do odpowiedniego klastra [26] lub dodatkowo dokonując *redukcji wymiarów* w obrębie grup i kodując przy pomocy wektora zmiennych rzeczywistych [97]. Jeszcze inne podejście stosuje macierze binarne do kodowania informacji o *rozmaitości*, tj. modele bazujące na *maszynach Boltzmanna* [153].

3. Ograniczeniem zaproponowanego modelu jest wspólna przestrzeń niskowymiarowa dla wszystkich stopni swobody w konfiguracji człowieka. Powoduje to, że model dopasowuje się do jednego charakterystycznego rodzaju ruchu, na podstawie którego jest nauczony, np. chodzenie, bieganie. Jednym ze sposobów na zwiększenie elastyczności modelu może być wprowadzenie dodatkowego *ukrytego modelu Markowa* (ang. hidden Markov model) odpowiadającego za przełączenie pomiędzy *rozmaitościami* odpowiadającymi różnym rodzajom ruchu. Innym sposobem jest podział stopni swobody *wektora stanu* na podzbiory lub hierarchie i modelowanie niskowymiarowych reprezentacji dla odpowiednich podgrup [91].

Rozdział 7

Badania empiryczne

Rozdział zawiera badania empiryczne weryfikujące jakość zaproponowanych w pracy metod. W szczególności przedstawione zostały wyniki *śledzenia ruchu człowieka* z wykorzystaniem *filtra cząsteczkowego* uwzględniającego niskowymiarową *rozmaitość*, który został opisany w rozdziale 4.3 oraz z wykorzystaniem *modelu wiarygodności* opartego na lokalnych deskryptorach, przedstawionego w rozdziale 5.4. Metody zostały porównane z technikami znanymi z literatury.

7.1 Zbiór danych

Do przeprowadzenia badań empirycznych, które zweryfikują jakość *śledzenia ruchu* z wykorzystaniem zaproponowanych metod konieczne jest porównanie otrzymanych estymat *wektora stanu* z rzeczywistymi konfiguracjami. Wymaga to posiadania specjalnie przygotowanych sekwencji ruchu, gdzie obrazy z kamer zostały zsynchronizowane z pomiarami z *systemu MOCAP*. Wtedy jakość metod może być weryfikowana w oparciu o sekwencje wideo i porównywana z rzeczywistymi ułożeniami ciała zarejestrowanymi przez *system MOCAP*. Zgodnie z wiedzą autora, na dzień dzisiejszy jedynym powszechnie dostępnym zbiorem benchmarkowym, który zawiera tak przygotowane sekwencje, jest zbiór *HumanEva* [140] i jest on obecnie standardowym narzędziem do testowania algorytmów do trójwymiarowej *estymacji pozy* i *śledzenia ruchu człowieka* [16, 19, 24, 34, 41, 59, 63, 94, 97, 108, 120, 140, 142, 128, 153].

Zbiór składa się z dwóch części – *HumanEva I* i *HumanEva II*, z których pierwsza zawiera sekwencje nagrane dla czterech różnych osób, a druga dla dwóch. Dla każdej postaci zarejestrowano różne rodzaje ruchu, jak chodzenie, bieganie, gestykulacja, boksowanie itp. Dodatkowo w przypadku pierwszej części zbioru każdy rodzaj ruchu został podzielony na kilka sekwencji, z czego niektóre z nich zawierają zsynchronizowany obraz i dane *MOCAP*, inne sam obraz lub same dane *MOCAP*.

Lp	Sekwencja	Ciąg	Liczba klatek	Fragment HumanEva
1	S1-Walk	Treningowy	350	S1, Walking, Train, Trial 1, 591-940
		Walidacyjny	300	S1, Walking, Train, Trial 3, 201-500
		Testowy	200	S1, Walking, Validate, Trial 1, 7-206
2	S1-Jog	Treningowy	220	S1, Jog, Train, Trial 1, 521-740
		Walidacyjny	200	S1, Jog, Train, Trial 3, 6-205
		Testowy	200	S1, Jog, Validate, Trial 1, 7-206
3	S2-Walk	Treningowy	350	S2, Walking, Train, Trial 1, 439-788
		Walidacyjny	300	S2, Walking, Train, Trial 3, 5-304
		Testowy	200	S2, Walking, Validate, Trial 1, 7-206
4	S2-Jog	Treningowy	350	S2, Jog, Train, Trial 1, 399-748
		Walidacyjny	300	S2, Jog, Train, Trial 3, 5-304
		Testowy	200	S2, Jog, Validate, Trial 1, 7-206
5	S3-Walk	Treningowy	350	S3, Walking, Train, Trial 3, 11-360
		Walidacyjny	300	S3, Walking, Train, Trial 3, 501-800
		Testowy	200	S3, Walking, Validate, Trial 1, 7-206
6	S3-Jog	Treningowy	350	S3, Jog, Train, Trial 1, 402-751
		Walidacyjny	300	S3, Jog, Train, Trial 3, 601-900
		Testowy	200	S3, Jog, Validate, Trial 1, 7-206

Tabela 7.1: Wyodrębnione sekwencje ruchu na potrzeby badań empirycznych

Zbiór *HumanEva I* zawiera obserwacje z siedmiu kamer – trzech kolorowych i czterech czarno białych, natomiast *HumanEva II* z czterech kolorowych kamer wysokiej jakości. Ponadto oba zbiory zawierają parametry wewnętrzne kamer, macierze rotacji i wektory prze-

sunięć otrzymane w procesie *kalibracji kamer* opisanym w rozdziale 2.2.1.

Na potrzeby badań empirycznych wykorzystany został zbiór *HumanEva I*. Wyodrębniono z niego sześć podzbiorów, z których każdy może być traktowany jako osobny zestaw benchmarkowy. Są to sekwencje zawierające chodzenie (Walk) i bieganie (Jog), odpowiednio dla postaci S1, S2 i S3[1]. Każdy z podzbiorów został podzielony na części treningową i testową zawierającą zsynchronizowany obraz i dane *MOCAP* oraz część walidacyjną zawierającą tylko dane *MOCAP*. Dokładny wykaz użytych sekwencji wraz z podanymi fragmentami zbioru *HuamEva I*, które zostały wykorzystane do ich zdefiniowania, został zamieszony w tabeli 7.1. Warto zwrócić uwagę, że użyto następującego schematu do konstrukcji ciągów: ciąg treningowy zawiera 350 przykładów, walidacyjny 300, a testowy 200 oraz ciągi treningowy i walidacyjny są wybrane z innych sekwencji (trial), aby na etapie uczenia przeciwdziałać zbytniemu dopasowaniu do danych (ang. overfitting). Wyjątki stanowią podzbiór S1-Jog, gdzie ciągi są krótsze oraz podzbiór S3-Walk, gdzie ciągi treningowy i walidacyjny wybrane zostały z ten samej sekwencji. Wynika to z faktu, że znaczna część danych w zbiorze *HumanEva* dla tych fragmentów jest uszkodzona i musiała zostać pominięta.

7.2 Badanie jakości śledzenia ruchu z użyciem filtra cząsteczkowego uwzględniającego niskowymiarową rozmaitość

Cel badania

Celem badania jest ocena jakości działania *filtra cząsteczkowego* uwzględniającego strukturę *rozmaitości* (MPF) opisanego w rozdziale 4.3 i porównanie go z metodami znanymi z literatury.

Metodyka badań

1. *Wstępne przetwarzanie danych*

   Niektóre sekwencje charakteryzowały się uszkodzeniem polegającym na gwałtow-

[1] Są to oznaczenia stosowane w zbiorze *HumanEva* do numerowania kolejnych osób.

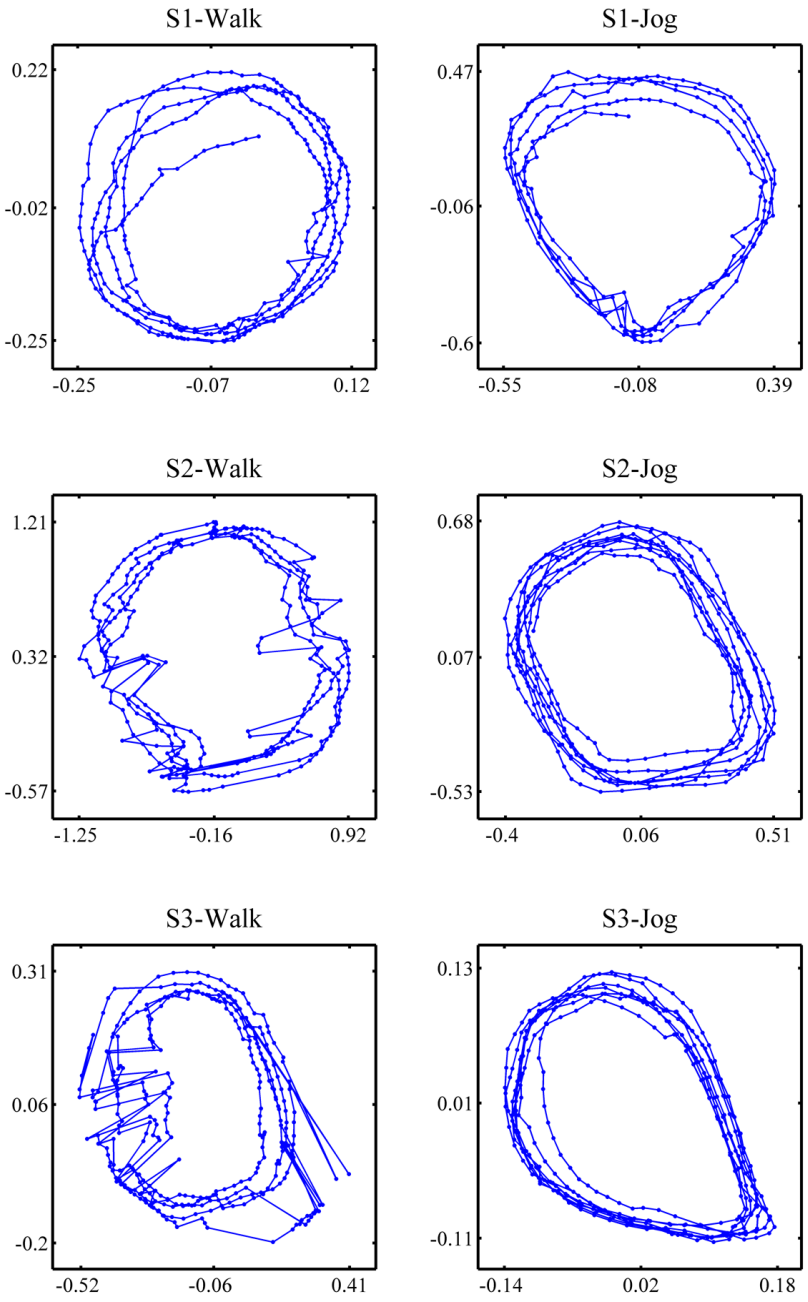
nych obrotach wokół lokalnej osi z dla obu ud oraz obu przedramion w sąsiednich klatkach obrazu, co powodowało spadek podobieństwa między kolejnymi *wektorami stanu*, a w konsekwencji problemy z prawidłowym odtworzeniem struktury *rozmaitości*. W przypadku przedramion obrót wokół osi z został wyzerowany, natomiast w przypadku ud obrót najpierw dodano do obrotu łydki, a następnie wyzerowano.

2. *Odtworzenie struktury rozmaitości*
   Do odtworzenia struktur rozmaitości dla poszczególnych sekwencji wykorzystano ciągi treningowe opisane w tabeli 7.1, które przetransformowano tak, by obroty były wyrażone przez zredukowane *kwaterniony*, a następnie przeskalowano przez stały współczynnik 10^2. Ustalono parametr $\gamma_x = 10^{-4}$ w funkcji jądra (6.33) oraz wymiar niskowymiarowej reprezentacji $\dim(\mathbf{z}) = 2$. Pozostałe parametry zostały wyznaczone w procesie optymalizacji. Rysunek 7.1 przedstawia niskowymiarowe reprezentacje $\mathbf{Z}$ ciągów treningowych uzyskane w procesie optymalizacji funkcji (6.25).

3. *Kalibracja parametrów modeli dynamiki*
   Parametry $\boldsymbol{\sigma}^2_{\mathbf{u}}$, $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}$, $\boldsymbol{\sigma}^2_{x\to z}$ i $\boldsymbol{\sigma}^2_{x\to x}$, odpowiednio w modelach (6.10), (6.11), (6.42) i (6.44), nauczono z użyciem estymatora postaci (6.6) z kwantylem rzędu 0.9. Dodatkowo dla wszystkich sekwencji chodzenia (Walk) wariancja została zwiększona 1.5 raza. W przypadku sekwencji biegania (Jog) zwiększono jedynie wariancję dla rotacji w globalnym układzie współrzędnych wokół osi z. Parametry $\boldsymbol{\sigma}^2_{z\to x}$ dla modelu (6.48) wyznaczono zgodnie z procedurą opisaną w rozdziale 6.2.3 przy użyciu estymatora największej wiarygodności i odpowiednich ciągów walidacyjnych.

4. *Model wiarygodności*
   Jako *model wiarygodności* wykorzystany został dwustronny model oparty na *sylwetkach* opisany równaniem (5.12). Parametry progowe w procesie *oddzielania tła* w zależności (5.10) zostały dobrane empirycznie dla każdej z kamer.

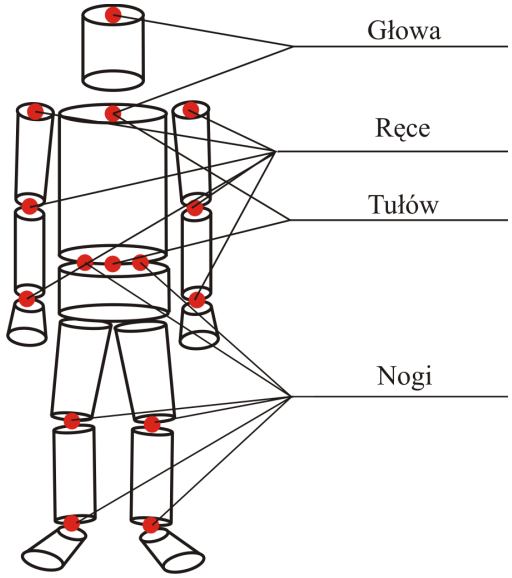
5. *Przeprowadzenie eksperymentu*
   Eksperyment został przeprowadzony z użyciem obrazów z trzech kolorowych kamer. Algorytm MPF został porównany z dwiema znanymi z literatury metodami: zwykłym *filtrem cząsteczkowy* (SIR), opisanym w rozdziale 4.1 i *wyżarzanym filtrem cząsteczkowym* (APF), opisanym w 4.2. Dla wszystkich sekwencji testowych (tabela 7.1) prze-



Rysunek 7.1: Zbiory punktów w niskowymiarowym układzie uzyskane na podstawie ciągów treningowych dla rozważanych sekwencji ruchu. Osie oznaczają odpowiednio współrzędne z^1 i z^2.

prowadzono po pięć niezależnych powtórzeń dla każdego z badanych algorytmów. W metodach MPF i SIR ustalono liczbę cząsteczek na 500, natomiast w metodzie APF zastosowano pięć warstw wyżarzania, po 100 cząsteczek na warstwę.

6. *Sposób oceny*
   Jakość *śledzenia ruchu* dla poszczególnych algorytmów została wyznaczona według miary opisanej w rozdziale 3.3. Punkty testowe zostały wyróżnione na *modelu ciała* opisanego w rozdziale 5.1. Są one wyznaczane dla rzeczywistego i wyestymowane-



Rysunek 7.2: Punkty testowe służące do oceny poprawności algorytmu śledzącego.

go *wektora stanu* z użyciem transformacji opisanych w rozdziale 2.1.1 i porównywane przy pomocy (3.28). Dokładne ułożenie punktów kontrolnych przedstawiono na rysunku 7.2. Jest to standardowe rozmieszczenie zaproponowane dla zbioru *HumanEva*. Dodatkowo, w celu bardziej szczegółowej analizy, punkty testowe zostały podzielone na cztery grupy: głowa, tułów, ręce i nogi.

7. *Weryfikacja statystyczna*
   Celem weryfikacji statystycznej jest pokazanie czy proponowana metoda (MPF) jest istotnie lepsza od pozostałych metod na zadanym poziomie istotności, bazując na

wynikach z sześciu wyróżnionych sekwencji. Jest to szczególny przypadek testowania wielu algorytmów na wielu zbiorach danych. Procedura testowania składa się z dwóch etapów:

- Najpierw wykorzystany zostanie jednostronny test Wilcoxona dla par obserwacji (ang. one-sided Wilcoxon signed-rank test) w celu wyliczenia p-wartości (ang. p-value) dla zbioru hipotez $H_i : m_{\mathrm{MPF}} \geqslant m_i$, gdzie m_{MPF} oznacza medianę błędów algorytmu MPF na różnych zbiorach danych, a m_i medianę błędów innego algorytmu. Test wypada korzystnie dla metody MPF, jeśli hipoteza zostaje odrzucona. Szczegóły testu Wilcoxona wraz z uzasadnieniem jego użycia do porównywania algorytmów zostały opisane w [43].
- Następnie hipotezy zostawiamy lub odrzucamy w oparciu o uzyskane p-wartości. Naiwne podejście sugerowałoby, że należy ustalić stały poziom istotności, np. 0.05 i odrzucać kolejne hipotezy, jeśli p-wartość jest poniżej tego poziomu. Należy jednak pamiętać, że poziom istotności oznacza prawdopodobieństwo popełnienia błędu pierwszego rodzaju, czyli odrzucenia hipotezy, gdy była ona prawdziwa. Kluczowe jest, aby ten błąd kontrolować. W przypadku testowania więcej niż jednej hipotezy, za każdym razem, gdy którąś z nich odrzucamy, jesteśmy narażeni na popełnienie błędu pierwszego rodzaju, a w konsekwencji prawdopodobieństwo, że odrzuciliśmy prawdziwą hipotezę w ciągu hipotez może być istotnie wyższe niż indywidualne poziomy istotności. Dlatego w teorii testowania wielu hipotez wprowadza się współczynnik FWER (ang. familywise error rate), który oznacza prawdopodobieństwo, że przynajmniej raz odrzucona została prawdziwa hipoteza i ten błąd chcemy kontrolować na ustalonym poziomie. W tym celu zastosować można procedurę Holma–Bonferroniego [70], która porządkuje otrzymane p-wartości w kolejności rosnącej $P_{(1)}, \ldots, P_{(m)}$, a następnie odrzuca te hipotezy $H_{(1)}, \ldots, H_{(k)}$, dla których:

$$P_{(k)} < \frac{\alpha}{m + 1 - k} \tag{7.1}$$

  oraz pozostawia hipotezy $H_{(k+1)}, \ldots, H_{(m)}$, jeśli $k + 1$ jest pierwszym indeksem, dla którego nie zachodzi warunek (7.1). Metoda ta gwarantuje, że $\mathrm{FWER} \leqslant \alpha$, gdzie α ustalamy na poziomie np. 0.05 i pozwala na „bezpieczne” porównanie

jednego algorytmu z wieloma innymi.

Uwagi

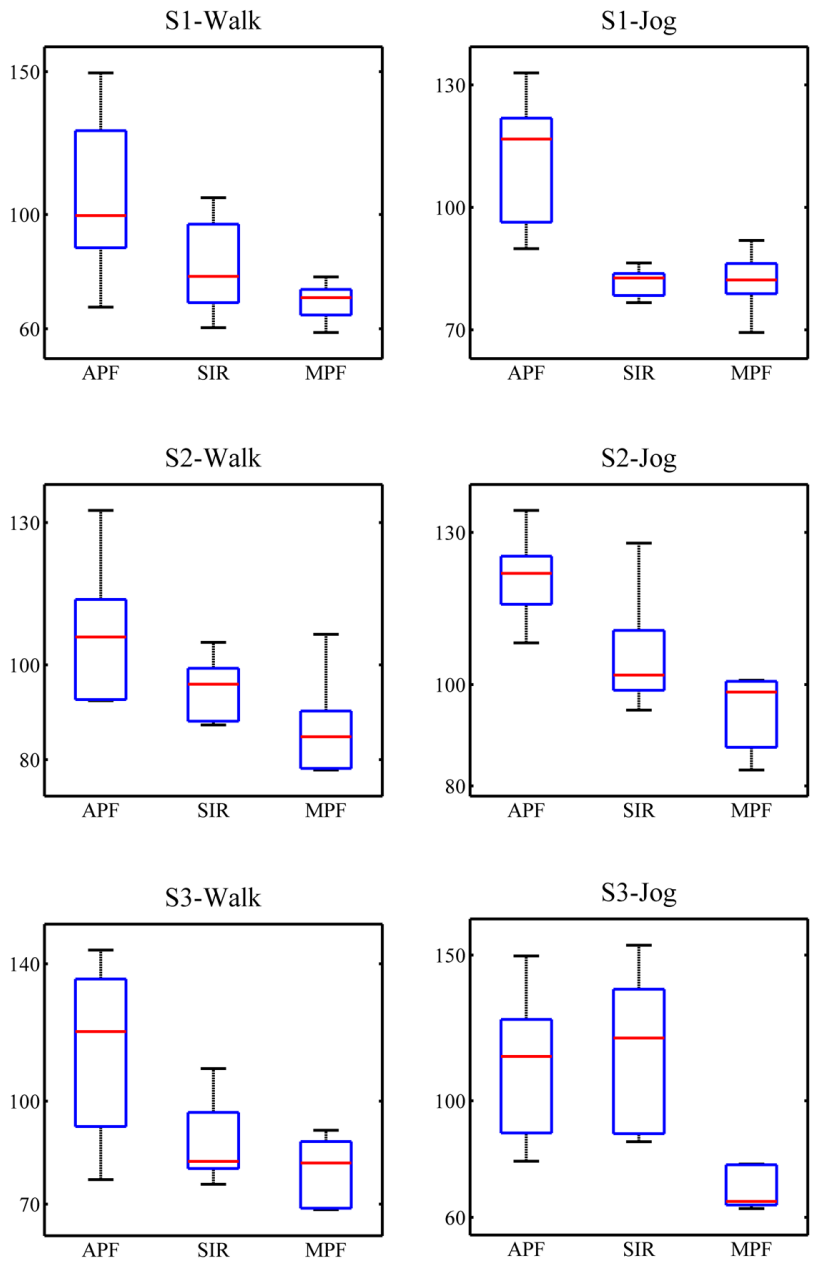
1. Badania symulacyjne zostały przeprowadzone w aplikacji HumanTracker stworzonej na potrzeby pracy doktorskiej. Program został napisany w języku C#.
2. Odtworzenie struktury *rozmaitości* wykonano przy pomocy pakietu FGPLVM [90] działającego w środowisku Matlab.

Wyniki

Szczegółowe błędy *śledzenia ruchu* dla wszystkich metod przedstawiono w tabeli 7.2. Rozkład błędów dla wszystkich sekwencji w postaci wykresów pudełkowych przedstawiono na rysunku 7.3. Przebieg uśrednionego błędu w poszczególnych klatkach zilustrowano na rysunku 7.4. W tabeli 7.3 przedstawiono średni błąd dla poszczególnych części ciała. Na rysunkach 7.5, 7.6, 7.7 zostały pokazane przykładowe klatki wraz z nałożonym *modelem ciała* wygenerowanym na podstawie wyników śledzenia dla poszczególnych metod. W tabeli 7.4 zostały przedstawione wyniki analizy statystycznej.

Dyskusja

Analizując wyniki przestawione w tabeli 7.2 i na wykresach na rysunku 7.3 widać, że algorytm MPF osiągnął lepsze rezultaty dla wszystkich sekwencji z wyjątkiem S1-Jog, gdzie był nieznacznie gorszy od zwykłego *filtra cząsteczkowego* (SIR). Przyczyną słabszych rezultatów dla sekwencji S2-Jog może być fakt, że w tym przypadku użyto krótszego ciągu treningowego, co zostało uzasadnione w rozdziale 7.1. Zdecydowanie najsłabiej zachowywał się *wyżarzany filtr cząsteczkowy* (APF), który we prawie we wszystkich sekwencjach był znacznie gorszy od pozostałych algorytmów. Przyczyną tego może być słaba jakość *sylwetek* otrzymywanych w procesie *oddzielania tła*, co powoduje wysoki poziom szumu w *modelu wiarygodności*, a w konsekwencji przesunięcie ekstremów rozkładu a posteriori, gdzie koncentrują się się cząsteczki w procesie wyżarzania. Metody SIR i MPF rozkładają cząsteczki



Rysunek 7.3: Rozkład błędów śledzenia (3.28) dla rozważanych sekwencji.

Sekwencja	Metoda	Iteracja					Śr.(Std.)
		1	2	3	4	5	
S1-Walk	APF	150	123	100	68	95	107(31)
	SIR	72	93	78	106	60	82(18)
	MPF	72	78	71	59	67	**69(7)**
S1-Jog	APF	90	98	118	133	117	111(17)
	SIR	79	83	77	86	83	**81(4)**
	MPF	82	92	82	84	69	82(8)
S2-Walk	APF	133	106	93	92	107	106(16)
	SIR	105	97	88	87	96	95(7)
	MPF	78	85	78	106	85	**86(12)**
S2-Jog	APF	118	108	134	122	122	121(9)
	SIR	128	105	102	100	95	106(13)
	MPF	89	98	101	83	101	**94(8)**
S3-Walk	APF	144	133	77	120	98	114(27)
	SIR	92	82	82	76	110	88(13)
	MPF	69	87	82	91	68	**79(10)**
S3-Jog	APF	92	79	115	121	150	111(27)
	SIR	133	90	86	153	121	117(29)
	MPF	63	78	65	78	65	**70(8)**

Tabela 7.2: Błąd śledzenia ruchu [mm] określony zależnością (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji.

bardziej równomiernie i przez to są w wyższym stopniu odporne na zaszumienie obserwacji.

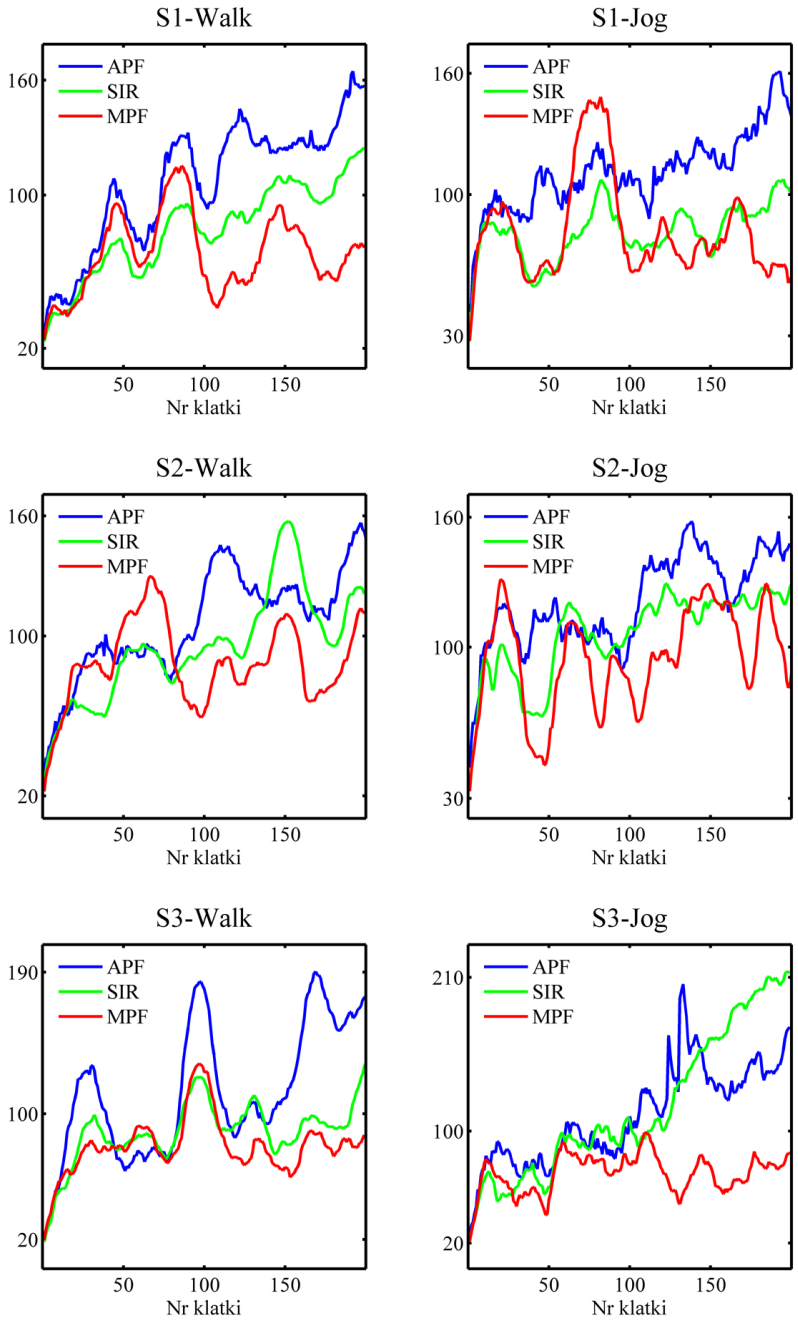
Szczegółowe wyniki dla poszczególnych części ciała (tabela 7.3) pokazują, że metoda MPF zawsze sprawdzała się lepiej dla tułowia i rąk. Wynika to z faktu, że te fragmenty ciała przesłaniają się nawzajem i nie mogą być rozróżnione na podstawie obserwacji w postaci binarnych *sylwetek*. Algorytm MPF korzysta z wiedzy apriorycznej zawartej w *rozmaitości* i dobiera ułożenia rąk i tułowia tak, by cała konfiguracja nie odchylała się od niej zbyt

Sekwencja	Metoda	Część ciała			
		Tułów	Głowa	Nogi	Ręce
S1-Walk	APF	50(5)	46(5)	110(34)	133(63)
	SIR	41(4)	**39(2)**	74(28)	111(32)
	MPF	**36(1)**	40(4)	**69(13)**	**85(7)**
S1-Jog	APF	44(6)	49(7)	96(21)	158(30)
	SIR	37(3)	**40(3)**	**63(8)**	122(17)
	MPF	**35(4)**	43(6)	80(19)	**106(7)**
S2-Walk	APF	73(6)	78(5)	120(35)	107(11)
	SIR	76(5)	76(3)	101(9)	98(11)
	MPF	**61(6)**	**72(4)**	**97(10)**	**85(20)**
S2-Jog	APF	66(9)	77(3)	106(12)	161(15)
	SIR	62(7)	**73(2)**	**86(8)**	146(40)
	MPF	**57(2)**	75(4)	102(9)	**101(11)**
S3-Walk	APF	48(3)	43(7)	143(35)	120(42)
	SIR	53(10)	44(10)	**93(17)**	104(14)
	MPF	**39(4)**	**35(3)**	99(26)	**81(11)**
S3-Jog	APF	50(13)	54(14)	105(15)	148(51)
	SIR	45(11)	51(11)	96(22)	172(48)
	MPF	**38(3)**	**40(3)**	**74(4)**	**81(15)**

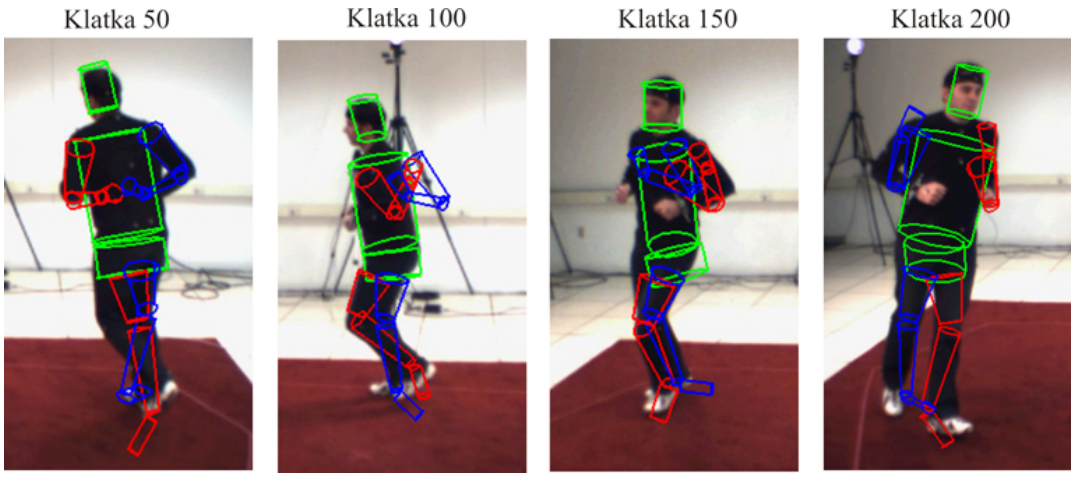
Tabela 7.3: Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm]. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach.

mocno. Problemy w działaniu metod APF i SIR dla wymienionych części ciała można zaobserwować na rysunkach 7.5 i 7.6, w odróżnieniu od metody MPF, która w tej sytuacji radzie sobie lepiej (rysunek 7.7). Z kolei algorytm MPF czasami gorzej estymuje położenie głowy niż zwykły *filtr cząsteczkowy*. Prawdopodobną przyczyną tego może być wykorzystanie modelu *GPLVM*, który jest za mało elastyczny i zbyt silnie wiąże ze sobą wszystkie stopnie swobody, wymuszając położenie głowy na podstawie położenia innych elementów.

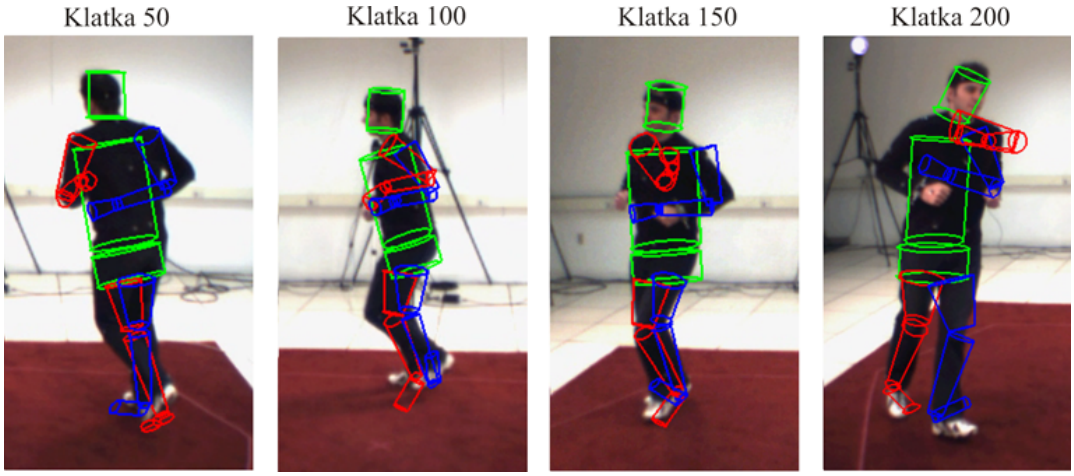
Osobną uwagę należy poświecić nogom, gdzie główną przyczyną rosnącego błędu jest



Rysunek 7.4: Przebieg średniego błędu w kolejnych klatkach śledzenia dla rozważanych sekwencji.

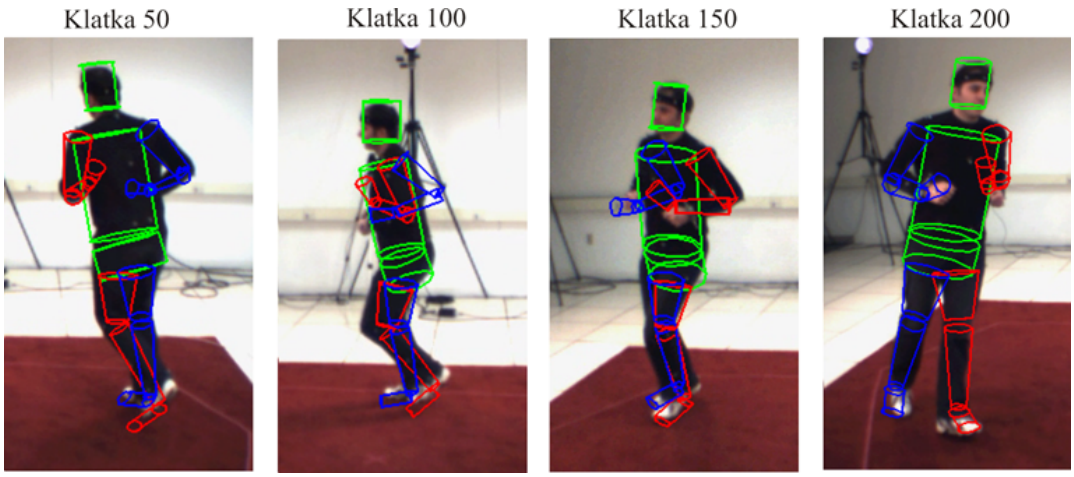


Rysunek 7.5: Przykładowy przebieg śledzenia dla wyżarzanego filtra cząsteczkowego (APF). Sekwencja S3-Jog. Kamera nr 2.



Rysunek 7.6: Przykładowy przebieg śledzenia dla zwykłego filtra cząsteczkowego (SIR). Sekwencja S3-Jog. Kamera nr 2.

zamiana nóg miejscami w procesie śledzenia. Efekt ten można zaobserwować dla klatki 200 na rysunku 7.6, gdzie prawa noga w *modelu ciała* ma niebieski kolor. Związany z tym problem jest bardzo trudny do wyeliminowania ponieważ nogi są lokalnie niemal nierozróżnialne. Co więcej algorytm MPF, który generuje konfiguracje ciała bliższe rzeczywistości niż SIR, jest często bardziej podatny na ten problem, co widać po wynikach w tabeli 7.3. W



Rysunek 7.7: Przykładowy przebieg śledzenia dla filtra cząsteczkowego uwzględniającego strukturę rozmaitości (MPF). Sekwencja S3-Jog. Kamera nr 2.

Metoda	Róznice w błędach z MPF						p-wartość	k	Wartość (7.1)	Odrzucamy H_i
	1	2	3	4	5	6				
APF	38	29	20	27	35	40	0.016	1	0.025	TAK
SIR	13	-1	9	12	9	47	0.031	2	0.05	TAK

Tabela 7.4: Różnice średnich błędów pomiędzy metodami z literatury i referencyjną metodą MPF dla rozważanych sekwencji testowych. Rezultaty testu statystycznego dla poziomu istotności $\alpha = 0.05$.

pracy [41] zaproponowano prostą heurystykę polegającą na losowej podmianie nóg, by złagodzić to zjawisko. Innym sposobem na zmniejszenie tego problemu może być rozszerzenie *modelu dynamiki* tak, by predykcja kolejnego stanu zależała od kilku stanów poprzednich.

Na wykresach na rysunku 7.4 przedstawiono przebieg uśrednionego błędu w kolejnych klatkach obrazu. Można zaobserwować, że algorytm MPF charakteryzuje się większą stabilnością niż APF i SIR, tj. błąd śledzenia od pewnego momentu zaczyna oscylować wokół pewnej stałej wartości i nie posiada istotnego trendu wzrostowego. Jest to bardzo korzystne zjawisko ponieważ zmniejsza szanse systemu śledzącego na zgubienie się, czyli osiągnięcie takiego poziomu błędu, że praktycznie niemożliwy jest powrót do prawidłowego szacowa-

nia konfiguracji ciała.

Podsumowując, cel badania został osiągnięty. Pokazano, że metoda MPF daje istotnie lepsze rezultaty niż metody znane z literatury. Wyniki badań zostały zweryfikowane testem statystycznym, którego szczegóły zostały przedstawione w tabeli 7.4. Ponieważ zostały odrzucone obie hipotezy zakładające, że każda z metod znanych z literatury jest nie gorsza od metody przedstawionej w pracy, to możemy stwierdzić, że z prawdopodobieństwem 0.95 algorytm MPF jest lepszy od pozostałych technik.

7.3 Badanie jakości śledzenia ruchu z użyciem modelu wiarygodności opartego na lokalnych deskryptorach

Cel badania

Celem badania jest ocena jakości *modelu wiarygodności* opartego na lokalnych deskryptorach (LD) opisanego w rozdziale 5.4 działającego samodzielnie i połączeniu z innymi modelami oraz porównanie go z modelami znanymi z literatury. Ocena odbywa się na podstawie skuteczności *śledzenia ruchu*.

Metodyka badań

1. *Uczenie modeli wyglądu*

   Do nauczenia parametrów odpowiednich *modeli wyglądu* (5.32) dla poszczególnych części ciała wykorzystano ciągi treningowe opisane w tabeli 7.1, które zawierają zsynchronizowane obrazy z kamer i dane z *systemu MOCAP*[2]. Na ich podstawie wygenerowano deskryptory przedstawiające wygląd z różnych perspektyw i w różnych fazach ruchu ustalonych części ciała. Dodatkowo stworzono zbiory losowo wybranych fragmentów obrazu, na podstawie których określono przykłady negatywne potrzebne w procesie uczenia (5.32). Ponadto z wygenerowanych zbiorów przykładów pozytywnych i negatywnych wybrano część treningową i walidacyjną, gdzie pierwsza posłużyła do ustalenia parametrów *modeli wyglądu*, a druga do określenia parametru *regularyzacji* w (5.30) oraz precyzji jądra w (5.31).

[2]Wyjątkiem jest sekwencja S3-Walk, dla której konieczne było użycie danych z innego fragmentu zbioru, gdyż wyróżniony w tabeli 7.1 ciąg treningowy zawiera jedynie dane MOCAP i nie zawiera obrazów z kamer.

2. *Model dynamiki*
   Jako *model dynamiki* został wykorzystano podstawowy model oparty na dyfuzji gaussowskiej opisany równaniem (6.3). Jego parametry zostały nauczone na podstawie sekwencji treningowych.

3. *Przeprowadzenie eksperymentu*
   Eksperyment został przeprowadzony z użyciem zwykłego *filtra cząsteczkowego*, gdzie ustalono liczbę cząsteczek na 500. Wykorzystano obrazy z trzech kolorowych kamer. Do porównania wybrano następujące *modele wiarygodności*:

   - Model oparty na *sylwetkach* (S) opisany równaniem (5.11).
   - Model oparty na *sylwetkach* połączony z modelem opartym na krawędziach (S+E), gdzie ten ostatni opisuje równanie (5.24).
   - Model oparty na dwustronnych *sylwetkach* (BS) opisany równaniem (5.12).
   - Model oparty na lokalnych deskryptorach (LD) opisany równaniem (5.36).
   - Model oparty na dwustronnych *sylwetkach* połączony z modelem opartym na lokalnych deskryptorach (BS+LD).

   Dla wszystkich sekwencji testowych przeprowadzono po pięć niezależnych powtórzeń dla każdego z wymienionych powyżej *modeli wiarygodności*. Łączenie różnych modeli wykonano zgodnie z techniką opisaną w rozdziale 5.5, gdzie poszczególne wagi zostały ustalone na 1.

4. *Sposób oceny*
   Jakość *śledzenia ruchu* została wyznaczona z użyciem punktów testowych w identyczny sposób, jak w przypadku badania opisanego w rozdziale 7.2.

5. *Weryfikacja statystyczna*
   Celem weryfikacji statystycznej jest pokazanie czy model oparty na lokalnych deskryptorach połączony z modelem na dwustronnych sylwetkach jest istotnie lepszy od modeli znanych z literatury na zadanym poziomie istotności. Użyto zbioru hipotez $H_i : m_{\mathrm{BS+LD}} \geqslant m_i$, gdzie m_i oznaczają medianę błędów dla pozostałych modeli. Została wykorzystana metoda testowania opisana w rozdziale 7.2.

Uwagi

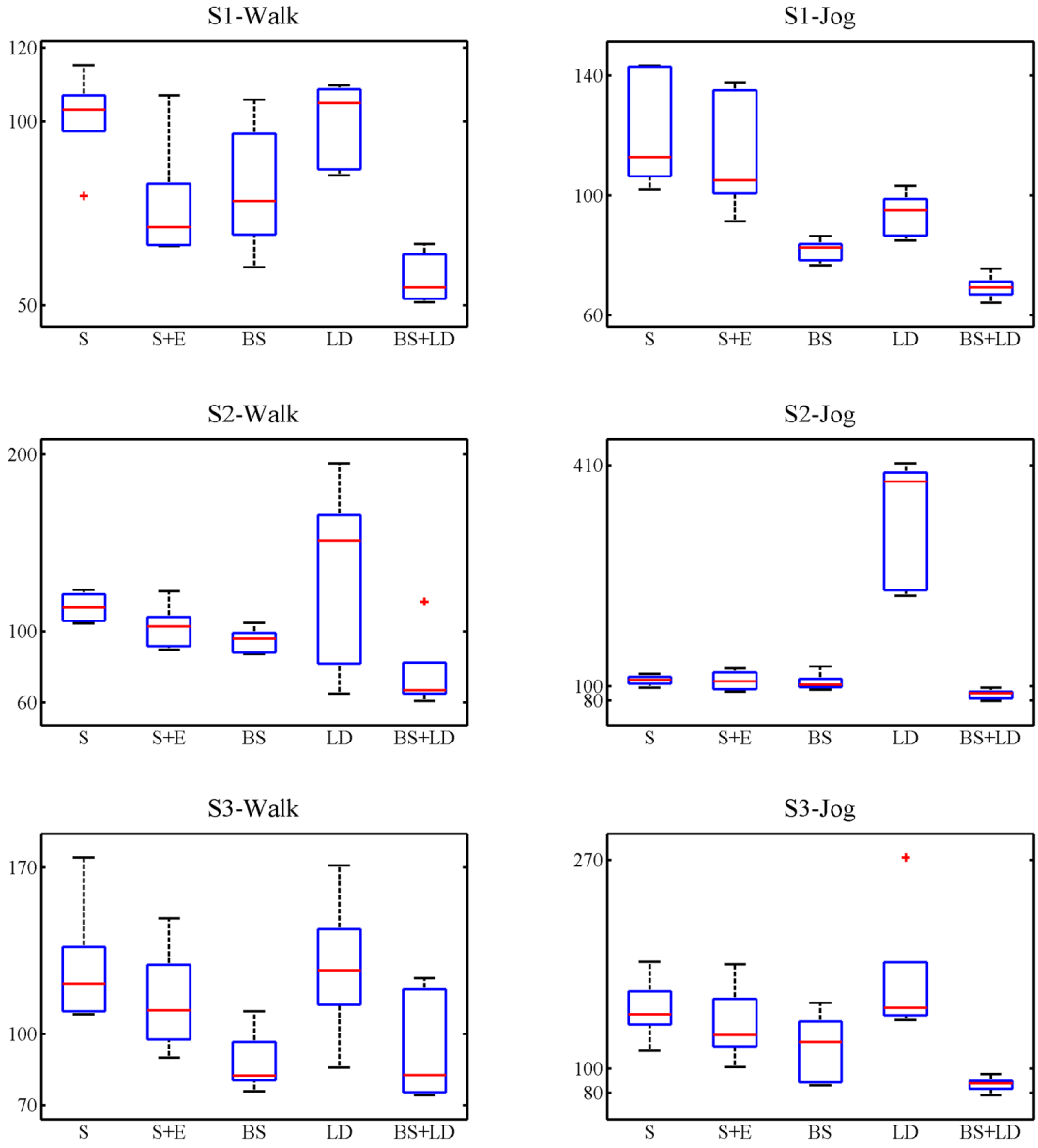
1. Badania symulacyjne zostały przeprowadzone w aplikacji HumanTracker stworzonej na potrzeby pracy doktorskiej.

2. Do uczenie *modeli wyglądu* wykorzystano implementację *Support Vector Machine* standardowo dołączoną do pakietu Matlab.

Wyniki

Sekwencja	Model	Iteracja					Śr.(Std.)
		1	2	3	4	5	
S1-Walk	S	103	115	103	80	104	101(13)
	S+E	66	71	75	107	66	77(17)
	BS	72	93	78	106	60	82(18)
	LD	85	110	108	87	105	99(11)
	BS+LD	67	55	52	63	51	**57(7)**
S2-Walk	S	123	120	104	113	107	114(8)
	S+E	92	90	103	103	123	102(13)
	BS	105	97	88	87	96	95(7)
	LD	65	151	195	88	156	131(53)
	BS+LD	117	67	71	61	66	**76(23)**
S3-Walk	S	108	174	124	121	110	128(27)
	S+E	90	122	100	110	149	114(23)
	BS	92	82	82	76	109	**88(13)**
	LD	121	135	171	127	86	128(31)
	BS+LD	123	74	76	83	117	95(24)

Tabela 7.5: Błąd śledzenia ruchu [mm] dla sekwencji z chodzeniem zadany przy pomocy zależności (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji.



Rysunek 7.8: Rozkład błędów śledzenia (3.28) dla rozważanych sekwencji i testowanych modeli wiarygodności.

Sekwencja	Model	Iteracja					Śr.(Std.)
		1	2	3	4	5	
S1-Jog	S	113	108	143	102	143	122(20)
	S+E	105	134	91	104	138	114(20)
	BS	79	83	77	86	83	81(4)
	LD	97	95	103	87	85	93(7)
	BS+LD	76	68	64	70	69	**69(4)**
S2-Jog	S	109	105	117	98	112	108(7)
	S+E	97	117	125	92	107	108(14)
	BS	128	105	102	100	95	106(13)
	LD	394	227	413	387	237	331(91)
	BS+LD	91	79	84	91	98	**88(7)**
S3-Jog	S	114	144	187	155	143	149(26)
	S+E	185	124	147	127	101	137(31)
	BS	133	90	86	153	121	117(29)
	LD	158	145	149	140	272	173(56)
	BS+LD	95	88	85	78	88	**87(6)**

Tabela 7.6: Błąd śledzenia ruchu [mm] dla sekwencji z bieganiem zadany przy pomocy zależności (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji.

Szczegółowe błędy *śledzenia ruchu* dla wszystkich modeli przedstawiono w tabelach 7.5 i 7.6, odpowiednio dla sekwencji z chodzeniem i bieganiem. Rozkład błędów dla wszystkich sekwencji w postaci wykresów pudełkowych przedstawiono na rysunku 7.8. Przebieg uśrednionego błędu w poszczególnych klatkach zilustrowano na rysunku 7.9. W tabelach 7.7 i 7.8 przedstawiono średni błąd dla poszczególnych części ciała, odpowiednio dla sekwencji z chodzeniem i bieganiem. Na rysunkach 7.10, 7.11, 7.12, 7.13, 7.14 zostały pokazane przykładowe klatki wraz z nałożonym *modelem ciała* wygenerowanym na podstawie wyników śledzenia dla poszczególnych modeli. W tabeli 7.9 zostały przedstawione wyniki analizy statystycznej.

Sekwencja	Model	Część ciała			
		Tułów	Głowa	Nogi	Ręce
S1-Walk	S	46(3)	38(3)	98(21)	134(12)
	S+E	46(3)	37(2)	78(47)	94(9)
	BS	41(4)	39(2)	74(28)	111(32)
	LD	49(12)	44(8)	128(9)	97(23)
	BS+LD	**32(3)**	**33(5)**	**56(19)**	**71(14)**
S2-Walk	S	78(6)	75(2)	110(14)	136(15)
	S+E	82(6)	77(1)	104(16)	112(20)
	BS	76(5)	76(3)	101(9)	98(11)
	LD	67(29)	58(24)	117(35)	179(90)
	BS+LD	**54(7)**	**51(5)**	**92(35)**	**72(22)**
S3-Walk	S	52(10)	40(12)	134(20)	161(56)
	S+E	57(8)	44(7)	118(24)	142(33)
	BS	53(9)	44(10)	**93(17)**	**104(14)**
	LD	71(17)	66(19)	142(38)	143(39)
	BS+LD	**39(7)**	**37(4)**	102(21)	116(46)

Tabela 7.7: Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm] dla sekwencji z chodzeniem. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach.

Dyskusja

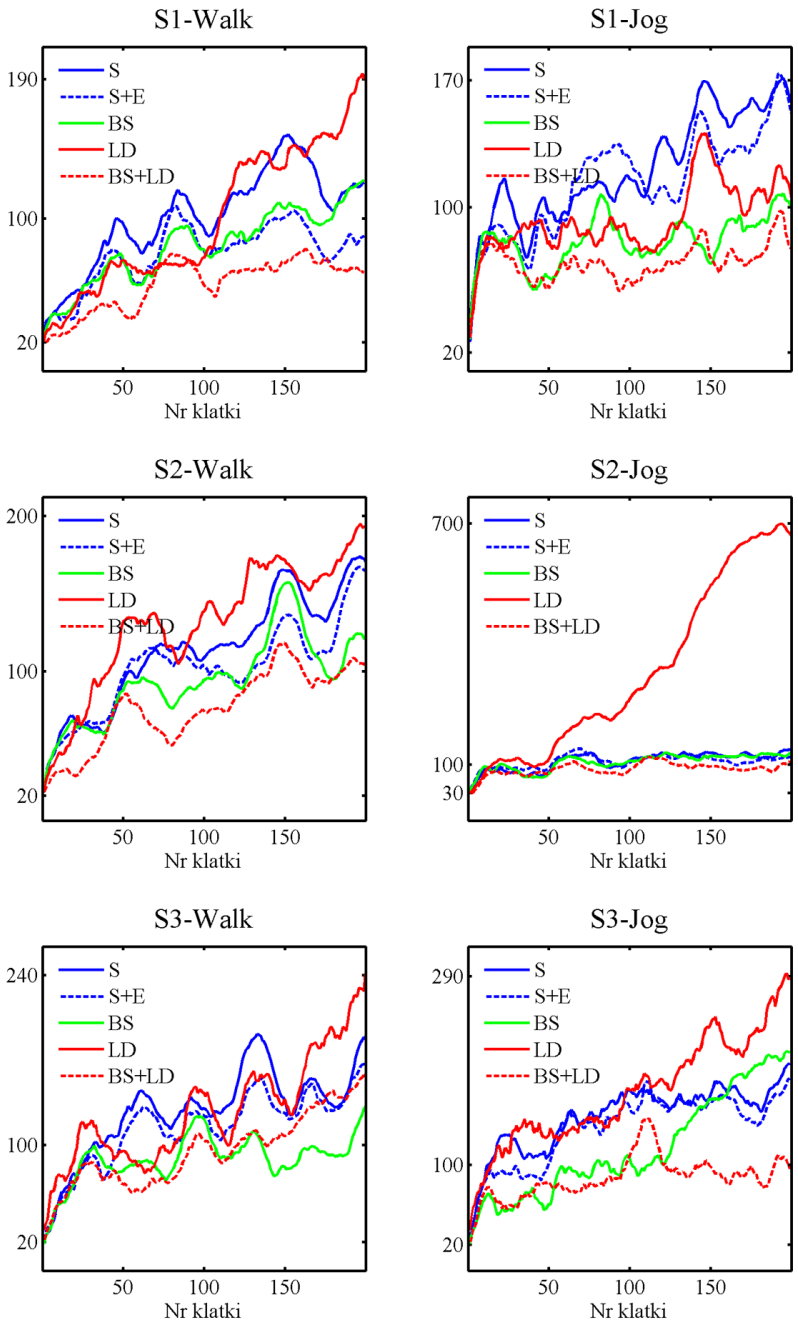
Analizując wyniki przestawione w tabelach 7.5 i 7.6 i na wykresach na rysunku 7.8 widzimy, że dla modelu BS+LD prawie zawsze dostajemy najlepsze rezultaty z wyjątkiem sekwencji S3-Walk, gdzie lepszy wynik został otrzymany przy zastosowaniu modelu BS. Może to wynikać z faktu, że dla tej sekwencji dysponowano gorszej jakości danymi z *systemu MOCAP*, co przełożyło się na jakość deskryptorów do uczenia *modelu wyglądu*. Korzystnym zjawiskiem jest również niska wariancja wyników otrzymanych dla modelu BS+LD, co sugeruje, że jest on odporny na losowość algorytmu śledzącego i przypadkowo nie zwraca

Sekwencja	Model	Część ciała			
		Tułów	Głowa	Nogi	Ręce
S1-Jog	S	40(4)	38(4)	141(33)	143(34)
	S+E	42(4)	43(4)	124(26)	141(29)
	BS	**37(3)**	40(3)	63(8)	122(17)
	LD	56(5)	44(5)	99(6)	110(19)
	BS+LD	43(2)	**38(4)**	**58(6)**	**95(7)**
S2-Jog	S	60(10)	71(4)	**79(7)**	159(13)
	S+E	70(11)	77(4)	88(20)	144(28)
	BS	62(7)	73(2)	86(8)	146(40)
	LD	251(127)	233(91)	323(115)	384(72)
	BS+LD	**49(10)**	**57(4)**	80(5)	**115(14)**
S3-Jog	S	47(7)	51(12)	143(27)	204(63)
	S+E	50(10)	51(15)	139(43)	178(69)
	BS	45(11)	51(11)	**96(22)**	172(48)
	LD	89(27)	66(17)	189(26)	205(117)
	BS+LD	**40(4)**	**40(7)**	97(13)	**100(22)**

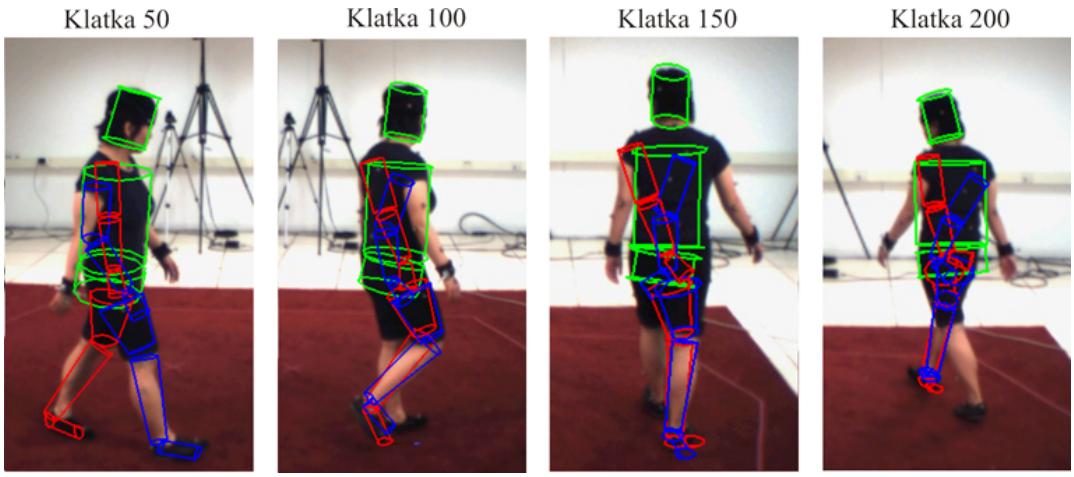
Tabela 7.8: Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm] dla sekwencji z bieganiem. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach.

wysokiej wiarygodności dla nieprawidłowych konfiguracji.

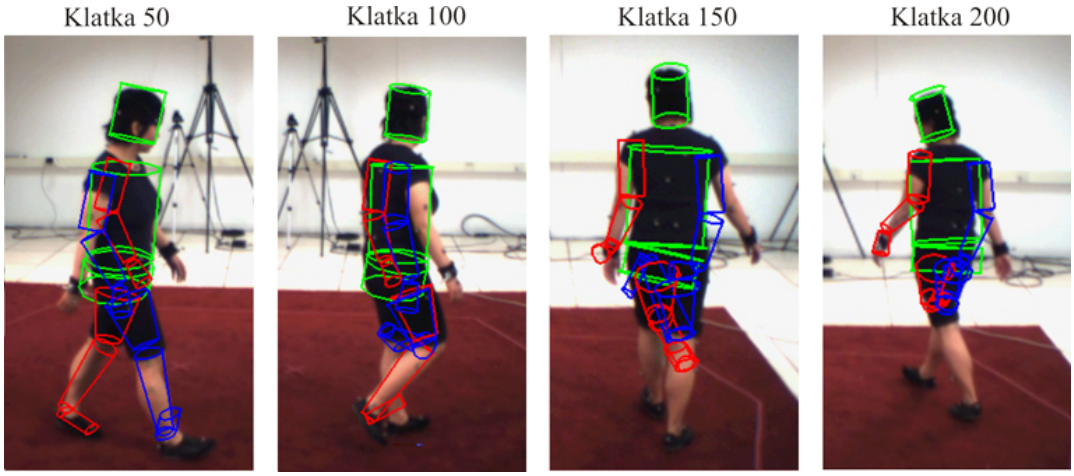
Standardowe modele stosowane w literaturze, tj. S i S+E, dają zdecydowanie gorsze wyniki od BS+LD, w szczególności różnica ta jest znacząca przy prawidłowej estymacji konfiguracji nóg i rąk, co można zaobserwować w tabelach 7.7 i 7.7 oraz na rysunkach 7.10 i 7.11. Wynika to przede wszystkim z faktu, że modele S i S+E bazują jedynie na informacji na ile pokryły się z obrazem, a nie na różnicy w pokryciach. Prowadzi to do istnienia nieprawidłowych konfiguracji, które mają wysoką wiarygodność, np. gdy obie nogi w *modelu ciała* dopasowują się do pojedynczej nogi na obrazie, co widać na rysunku 7.10.



Rysunek 7.9: Przebieg średniego błędu w kolejnych klatkach śledzenia dla rozważanych sekwencji i wybranych modeli wiarygodności.

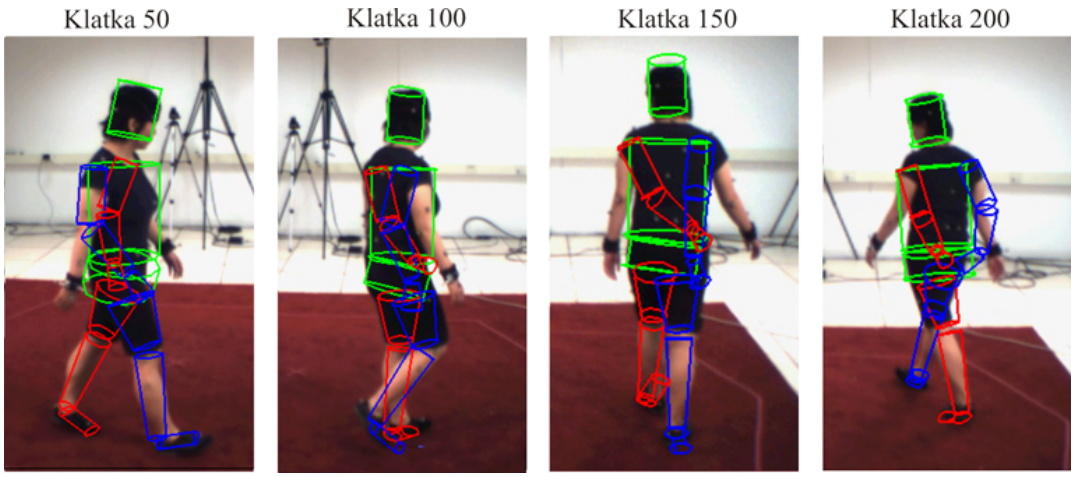


Rysunek 7.10: Przykładowy przebieg śledzenia dla modelu opartego na sylwetkach (S). Sekwencja S1-Walk. Kamera nr 2.

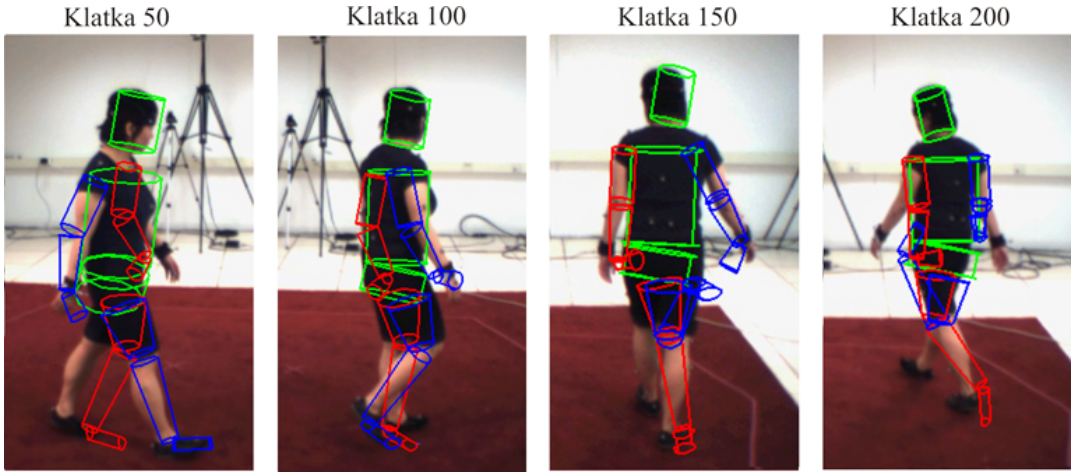


Rysunek 7.11: Przykładowy przebieg śledzenia dla modelu opartego na sylwetkach i krawędziach (S+E). Sekwencja S1-Walk. Kamera nr 2.

Nieco lepiej od S i S+E, zachowuje się model BS. Korzysta on z dwustronnej informacji, tj. na ile model dopasowuje się do obrazu i na ile obraz pokrył się z modelem. Dzięki temu przeważnie bardziej poprawnie estymowana jest konfiguracja nóg niż w przypadku S i S+E (tabele 7.8 i 7.7), co widać na rysunku 7.12. Jednakże w przypadku rąk, gdzie występuje duża strata informacji w momencie pokrycia z tułowiem, model nie sprawdza się (rysunek



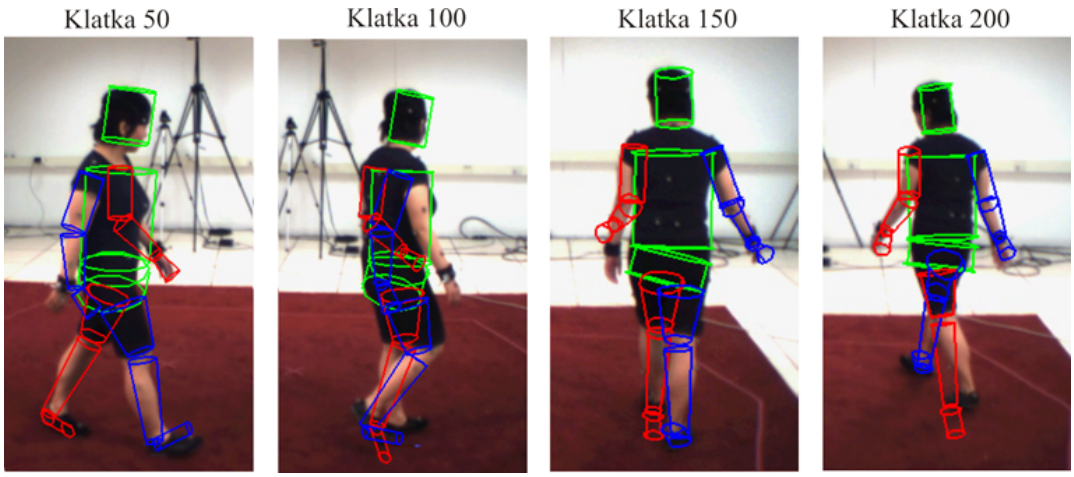
Rysunek 7.12: Przykładowy przebieg śledzenia dla modelu opartego na dwustronnych sylwetkach (BS). Sekwencja S1-Walk. Kamera nr 2.



Rysunek 7.13: Przykładowy przebieg śledzenia dla modelu opartego na lokalnych deskryptorach (LD). Sekwencja S1-Walk. Kamera nr 2.

7.12).

Warto zauważyć, że słabe wyniki daje samodzielne zastosowanie modelu opartego na lokalnych deskryptorach (LD). Najprawdopodobniej wynika to z faktu, że ocena dopasowania trójwymiarowego *modelu ciała* do obrazu na podstawie dopasowania niewielkiej liczby punktów jest niewystarczająca. Ponadto na wykresach na rysunku 7.9 widać, że dla



Rysunek 7.14: Przykładowy przebieg śledzenia dla modelu opartego na dwustronnych sylwetkach i lokalnych deskryptorach (BS+LD). Sekwencja S1-Walk. Kamera nr 2.

Model	Róznice w błędach z BS+LD						p-wartość	k	Wartość (7.1)	Odrzucamy H_i
	1	2	3	4	5	6				
S	44	53	38	20	33	62	0.0156	1	0.0167	TAK
S+E	20	45	26	20	19	50	0.0156	2	0.025	TAK
BS	25	12	19	18	-7	30	0.0313	3	0.05	TAK

Tabela 7.9: Różnice średnich błędów pomiędzy modelami znanymi z literatury i referencyjnym modelem BS+LD dla rozważanych sekwencji testowych. Rezultaty testu statystycznego dla poziomu istotności $\alpha = 0.05$.

większości sekwencji model LD od pewnego momentu zaczyna osiągać najwyższy błąd. W szczególności gwałtowny wzrost błędu można zauważyć dla sekwencji S2-Jog. Prowadzi to do wniosku, że system śledzący przestaje być stabilny i zaczyna gubić śledzoną postać, co częściowo widać na rysunku 7.13.

Połączenie modelu LD z modelem BS lokalnie uzupełnia niejednoznaczności wynikające ze stosowania binarnej *sylwetki*, „wprowadzając" w wybrane miejsca informację zawartą w *modelach wyglądu* dla wyróżnionych punktów. Powoduje to istotną poprawę jakości śledzenia, w szczególności dla kończyn górnych, co możemy zaobserwować w tabelach 7.7 i

7.8 oraz na rysunku 7.14.

Podsumowując zaprezentowane wyniki badań należy stwierdzić, że cel badania został osiągnięty. Pokazano, że model oparty na lokalnych deskryptorach połączony z modelem opartym dwustronnych *sylwetkach* (BS+LD) daje istotnie lepsze rezultaty niż modele znane z literatury. Wyniki badań zostały zweryfikowane testem statystycznym (tabela 7.9), dzięki któremu możemy stwierdzić, że z prawdopodobieństwem 0.95 model BS+LD jest lepszy od modeli S, S+E i BS. Dodatkowo pokazano, że model LD nie powinien być stosowany jako samodzielny *model wiarygodności*.

Rozdział 8

Uwagi końcowe

8.1 Oryginalny wkład pracy w dziedzinę śledzenia ruchu człowieka

Nowymi elementami przedstawionymi w pracy, które poszerzają dotychczasową wiedzę w zakresie bezznacznikowego *śledzenia ruchu człowieka*, są: opracowanie *filtra cząsteczkowego* uwzględniającego strukturę niskowymiarowej *rozmaitości*, opracowanie składowych *modeli dynamiki* na potrzeby zaproponowanego algorytmu śledzącego oraz opracowanie *modelu wiarygodności* opartego na lokalnych deskryptorach. Poniżej krótko omówiony został każdy z elementów.

Filtr cząsteczkowy uwzględniający niskowymiarową rozmaitość. W pracy wyprowadzono nowe zadanie *filtrowania* (rozdział 3.2.1), które zakłada, że bieżący *wektor stanu* wpływa jednocześnie na przyszły *wektor stanu* i jego niskowymiarową reprezentację, która koduje położenie na *rozmaitości*. Zaproponowano szczególny rodzaj *filtra cząsteczkowego*, który rozwiązuje w sposób przybliżony postawione zadanie *filtrowania* (rozdział 4.3). Opracowany algorytm jest niezależny od wyboru składowych *modelu dynamiki* i *modelu wiarygodności*, i stanowi ogólną procedurę do *śledzenia ruchu* z wiedzą aprioryczną zawartą w strukturze niskowymiarowej *rozmaitości*. Na podstawie wyników badań empirycznych pokazano, że zaproponowany algorytm jest istotnie lepszy od znanych z literatury metod (rozdział 7.2).

Składowe modelu dynamiki uwzględniające wiedzę o rozmaitości. W pracy zapro-

ponowano model dynamiki, który może być wykorzystany w opracowanym algorytmie śledzącym (rozdział 6.2). W szczególności podano postaci jego składowych, tj. *modelu dynamiki* po *rozmaitości* (rozdział 6.2.2) i *modelu dynamiki* w przestrzeni stanów z uwzględnieniem *rozmaitości* (rozdział 6.2.3). Przedstawiono techniki pozwalające na estymację ich parametrów.

Model wiarygodności oparty na lokalnych deskryptorach. W pracy opracowano nowy *model wiarygodności*, który pozwala wykorzystać wiedzę o lokalnym wyglądzie poszczególnych fragmentów ciała (rozdział 5.4). Dodatkowo zaproponowana została przykładowa postać deskryptorów (rozdział 5.4.1), opisujących wybrane punkty na ciele i sposób budowania *modelu wyglądu* na ich podstawie (rozdział 5.4.2). W oparciu o wyniki badań empirycznych pokazano, że zaproponowany model powinien być traktowany jako uzupełnienie brakującej informacji dla modelu opartego na *sylwetkach* (rozdział 7.3).

8.2 Kierunki dalszych badań

Analiza problemu *śledzenia ruchu człowieka* z użyciem podejścia *generującego* pozwoliły wskazać potencjalne kierunki dalszych badań. Poniżej wymieniono najważniejsze z nich.

1. Dekompozycja *wektora stanu* i zastosowanie *modelu opartego na częściach*. Opracowanie indywidualnych *modeli wiarygodności* i hierarchicznego rozkładu modelującego strukturę *rozmaitości*. Na tej podstawie opracowanie *filtra cząsteczkowego*, który pozwalałby generowanie cząsteczek, gdzie część wymiarów byłaby warunkowana innymi wymiarami. Wydaje się, że to jest kierunek, do stworzenia metod śledzących, które będą radziły sobie ze skomplikowanymi rodzajami ruchu. Przydane mogą być tutaj koncepcje zawarte w pracach [41, 131].

2. Opracowanie techniki łączenia *modeli wiarygodności* innej niż zwykłe ważenie. W szczególności zastosować można model *mieszaniny ekspertów* (ang. mixture of experts) [76], który wybierałby, która funkcja wiarygodności najlepiej nadaje się do porównania ustalonego obrazu i konfiguracji ciała.

3. Opracowanie dedykowanych deskryptorów opisujących lokalne fragmenty ciała. Zastosowanie mogą tutaj znaleźć *głębokie maszyny Boltzmanna* (ang. deep Boltzmann

machine) [12], które pozwalają na nauczenie z danych odpowiednich cech opisujących dany rodzaj obrazów, bez konieczności zadawania ich z góry, w odróżnieniu od używanych w pracy *falek Haara*.

4. Opracowanie *modelu wiarygodności*, który łączyłby zaproponowany *model wiarygodności* oparty na lokalnych deskryptorach z zaawansowanym detektorem poszczególnych części ciała, jak np. [172]. Ma to na celu otrzymanie informacji zwrotnej o prawdopodobnym położeniu części ciała na obrazie i wydaje się, że jest kluczem do stworzenia wysoce efektywnych *modeli wiarygodności*.

5. Przetestowanie bardziej elastycznych technik do modelowania struktury niskowymiarowej *rozmaitości*, jak np. hierarchiczny *GPLVM* [91], mieszanina analiz czynnikowych [61].

6. Opracowanie *modelu dynamiki*, który w razie potrzeby pozwalałby uwzględnić wiedzę z więcej niż jednego stanu wstecz. Wydaje się, że to jest droga do wyeliminowania problemu z zamieniającymi się nogami podczas śledzenia. Przydatna może być tutaj koncepcja modelu *Variable Length Markov Model* [26].

Bibliografia

[1] Abdi H., Williams L.J., *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics, 2(4):433—459, 2010. [cytowanie na str. 90]

[2] Agarwal A., Daumé III H., Gerber S., *Learning multiple tasks using manifold regularization*, in *NIPS '10 Proceedings of the Advances in Neural Information Processing Systems*, 2010. [cytowanie na str. 95, 97]

[3] Agarwal A., Triggs B., *Hyperfeatures – multilevel local coding for visual recognition*, in *ECCV '06 Proceedings of the 9th European conference on Computer Vision - Volume Part I*, pp. 30–43, 2006. [cytowanie na str. 3, 76]

[4] Agarwal A., Triggs B., *Recovering 3D human pose from monocular images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):44–58, 2006. [cytowanie na str. 4]

[5] Andriluka M., Roth S., Schiele B., *People-tracking-by-detection and people-detection-by-tracking*, in *CVPR '08 Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. [cytowanie na str. 8, 82]

[6] Andriluka M., Roth S., Schiele B., *Pictorial structures revisited: People detection and articulated pose estimation*, in *CVPR '09 Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [cytowanie na str. 8, 82]

[7] Andriluka M., Roth S., Schiele B., *Monocular 3D pose estimation and tracking by detection*, in *CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [cytowanie na str. 8, 82]

[8] Baker A., *Matrix Groups: An Introduction to Lie Group Theory*, Springer-Verlag, 2002. [cytowanie na str. 16]

[9] Balan A.O., Black M.J., *An adaptive appearance model approach for model-based articulated object tracking*, in *CVPR '06 Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 758–765, 2006. [cytowanie na str. 61, 81]

[10] Balan A.O., Sigal L., Black M.J., *A quantitative evaluation of video-based 3D person tracking*, in *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance. Evaluation of Tracking and Surveillance*, pp. 349–356, 2005. [cytowanie na str. 41]

[11] Belkin M., Niyogi P., *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation, 15(6):1373–1396, 2003. [cytowanie na str. 3, 94]

[12] Bengio Y., *Learning deep architectures for AI*, Foundations and Trends in Machine Learning, 2(1):1–127, 2009. [cytowanie na str. 3, 76, 126]

[13] Bengio Y., Paiement J.F., Vincent P., Delalleau O., Le Roux N., Ouimet M., *Out-of-Sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering*, in *NIPS '03 Proceedings of the Advances in Neural Information Processing Systems*, pp. 177–184, 2003. [cytowanie na str. 95]

[14] Berg A.C., Malik J., *Geometric blur for template matching*, in *CVPR '01 Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 607–614, 2001. [cytowanie na str. 3, 76]

[15] Berger J.O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985. [cytowanie na str. 32]

[16] Bergtholdt M., Kappes J., Schmidt S., Schnörr C., *A study of parts-based object class detection using complete graphs*, International Journal of Computer Vision, 87:93—117, 2010. [cytowanie na str. 8, 41, 82, 98]

[17] Bishop C.M., *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006. [cytowanie na str. 3, 25, 32, 36, 37, 39, 43, 79, 92]

[18] Bishop C.M., Svensén M., Williams C.K.I., *GTM: The generative topographic mapping*, Neural Computation, 10(1):215—-234, 1998. [cytowanie na str. 3, 94]

[19] Bo L., Sminchisescu C., *Twin gaussian processes for structured prediction*, International Journal of Computer Vision, 87:28—-52, 2010. [cytowanie na str. 4, 5, 23, 38, 41, 98]

[20] Bodor R., Drenner A., Fehr D., Masoud O., Papanikolopoulos N., *View-independent human motion classification using image-based reconstruction*, Image and Vision Computing, 27:1194––1206, 2009. [cytowanie na str. 4]

[21] Boyd S., Vandenberghe L., *Convex Optimization*, Cambridge University Press, Cambridge, 2004. [cytowanie na str. 78]

[22] Boyd S.P., Barratt C.H., *Linear Controller Design: Limits of Performance*, Prentice-Hall, 1991. [cytowanie na str. 12]

[23] Breiman L., *Random forests*, Machine Learning, 45(1):5–32, 2001. [cytowanie na str. 79]

[24] Brubaker M.A., Fleet D.J., Hertzmann A., *Physics-based person tracking using the anthropomorphic walker*, International Journal of Computer Vision, 87:140—-155, 2010. [cytowanie na str. 6, 7, 23, 41, 43, 98]

[25] Bubnicki Z., *Modern Control Theory*, Springer-Verlag, Berlin, 2005. [cytowanie na str. 12]

[26] Caillete F., Galata A., Howard T., *Real-time 3-D human body tracking using learnt models of behaviour*, Computer Vision and Image Understanding, 109:112—-125, 2008. [cytowanie na str. 6, 7, 23, 43, 61, 97, 126]

[27] Canny J., *A computational approach to edge detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986. [cytowanie na str. 69]

[28] Cao D., Masoud O.T., Boley D., Papanikolopoulos N., *Human motion recognition using support vector machines*, Computer Vision and Image Understanding, 113:1064—-1075, 2009. [cytowanie na str. 4]

[29] Caprile B., Torre V., *Using vanishing points for camera calibration*, International Journal of Computer Vision, 4(2):127–139, 1990. [cytowanie na str. 27]

[30] Casella G., Robert C.P., *Rao-blackwellisation of sampling schemes*, Biometrika, 83(1):81–94, 1996. [cytowanie na str. 50]

[31] Cevher V., Sankaranarayanan A., Duarte M.F., Reddy D., Baraniuk R.G., Chellappa R., *Compressive sensing for background subtraction*, in *ECCV '08 Proceedings of the 10th European Conference on Computer Vision: Part II*, pp. 155–168, 2008. [cytowanie na str. 3, 64]

[32] Chang I., Lin S., *3D human motion tracking based on a progressive particle filter*, Pattern Recognition, 43(10):3621–3635, 2010. [cytowanie na str. 7, 43, 61, 66, 72, 82]

[33] Chen J., Kim M., Wang Y., Ji Q., *Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition*, in *CVPR '09 Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [cytowanie na str. 38, 66, 94]

[34] Cheng S.Y., Trivedi M.M., *Articulated human body pose inference from voxel data using a kinematically constrained gaussian mixture model*, in *CVPR '07 Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation*, 2007. [cytowanie na str. 8, 23, 41, 61, 98]

[35] Cherkassky V., Mulier F., *Learning From Data. Concepts, Theory, and Methods*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007. [cytowanie na str. 32]

[36] Chopin N., *Central limit theorem for sequential monte carlo methods and its application to bayesian inference*, Annals of Statistics, 32(6):2385–2411, 2004. [cytowanie na str. 44]

[37] Cremers D., Rousson M., Deriche R., *A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape*, International Journal of Computer Vision, 72(2):195–215, 2007. [cytowanie na str. 3]

[38] Crisan D., Doucet A., *A survey of convergence results on particle filtering methods for practitioners*, IEEE Transactions on Signal Processing, 50(3):736–746, 2002. [cytowanie na str. 44]

[39] Dalal N., Triggs B., *Histograms of oriented gradients for human detection*, in *CVPR '05 Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005. [cytowanie na str. 3, 76]

[40] Daubney B., Gibson D., Campbell N., *Real-time pose estimation of articulated objects using low-level motion*, in *CVPR '08 Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. [cytowanie na str. 8, 81, 82]

[41] Daubney B., Xie X., *Tracking 3D human pose with large root node uncertainty*, in *CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [cytowanie na str. 7, 15, 41, 43, 53, 61, 66, 82, 98, 111, 125]

[42] Dempster A.P., Laird N.M., Rubin D.B., *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society B, 39(1):1—-38, 1977. [cytowanie na str. 3]

[43] Demšar J., *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, 7:1–30, 2006. [cytowanie na str. 104]

[44] Deutscher J., Reid I., *Articulated body motion capture by stochastic search*, International Journal of Computer Vision, 61(2):185–205, 2005. [cytowanie na str. 6, 7, 23, 43, 50, 51, 52, 53, 61, 66, 72, 82, 83]

[45] Doucet A., de Freitas N., Murphy K.P., Russell S.J., *Rao-blackwellised particle filtering for dynamic bayesian networks*, in *UAI '00 Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 176–183, 2000. [cytowanie na str. 50]

[46] Doucet A., Johansen A.M., *A tutorial on particle filtering and smoothing: Fifteen years later*, in *Oxford Handbook of Nonlinear Filtering*, Oxford University Press, 2009. [cytowanie na str. 37, 44, 49, 51]

[47] Eichner M., Ferrari V., *Better appearance models for pictorial structures*, in *BMVC '09 Proceedings of the British Machine Vision Conference*, 2009. [cytowanie na str. 8, 82]

[48] Ek C.H., Torr P.H.S., Lawrence N.D., *Gaussian process latent variable models for human pose estimation*, in *MLMI'07 Proceedings of the 4th international conference on Machine learning for multimodal interaction*, pp. 132–143, 2007. [cytowanie na str. 4, 5, 38, 93, 94]

[49] Elgammal A.M., Harwood D., Davis L.S., *Non-parametric model for background subtraction*, in *ECCV '00 Proceedings of the 6th European Conference on Computer Vision-Part II*, pp. 751–767, 2000. [cytowanie na str. 3, 64]

[50] Erol A., Bebis G., Nicolescu M., Boyle R.D., Twombly X., *Vision-based hand pose estimation: A review*, Computer Vision and Image Understanding, 108:52—73, 2007. [cytowanie na str. 4]

[51] Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D., *Object detection with discriminatively trained part-based models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010. [cytowanie na str. 3, 7]

[52] Felzenszwalb P.F., Huttenlocher D.P., *Pictorial structures for object recognition*, International Journal of Computer Vision, 61(1):55–79, 2005. [cytowanie na str. 3, 7]

[53] Felzenszwalb P.F., Huttenlocher D.P., *Distance transforms of sampled functions*, Theory of Computing, 8:415–428, 2012. [cytowanie na str. 8]

[54] Fischler M.A., Elschlager R.A., *The representation and matching of pictorial structures*, IEEE Transactions on Computers, 22(1):67–92, 1973. [cytowanie na str. 7]

[55] Forsyth D.A., Ponce J., *Computer Vision - A Modern Approach*, Prentice-Hall, Upper Saddle River, NJ, 2002. [cytowanie na str. 3, 25, 26, 69]

[56] Freifeld O., Weiss A., Zuffi S., Black M.J., *Contour people: A parameterized model of 2D articulated human shape*, in *CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [cytowanie na str. 61]

[57] Freund Y., Schapire R.E., *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences, 55(1):119—139, 1997. [cytowanie na str. 79]

[58] Gall J., de Aguiar C.S.E., Theobalt C., Rosenhahn B., Seidel H.P., *Motion capture using joint skeleton tracking and surface estimation*, in *CVPR '09 Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [cytowanie na str. 23, 61, 72]

[59] Gall J., Rosenhahn B., Brox T., Seidel H.P., *Optimization and filtering for human motion capture. a multi-layer framework*, International Journal of Computer Vision, 87:75—-92, 2010. [cytowanie na str. 6, 41, 61, 72, 98]

[60] Gerber S., Tasdizen T., Whitaker R., *Dimensionality reduction and principal surfaces via Kernel Map Manifolds*, in *ICCV '09 Proceedings of the 12th IEEE International Conference on Computer Vision*, pp. 529–536, 2009. [cytowanie na str. 95, 97]

[61] Ghahramani Z., Beal M.J., *Variational inference for bayesian mixtures of factor analysers*, in *NIPS '99 Proceedings of the Advances in Neural Information Processing Systems*, pp. 449–455, 1999. [cytowanie na str. 3, 126]

[62] Goldstein H., *Classical mechanics*, Addison-Wesley, 1980. [cytowanie na str. 13]

[63] Gonczarek A., Tomczak J.M., *Manifold regularized particle filter for articulated human motion tracking*, in *ICSS '13 Proceedings of the 2013 International Conference on Systems Science (to appear)*, 2013. [cytowanie na str. 23, 38, 41, 61, 94, 98]

[64] Gonczarek A., Tomczak J.M., Świątek J., *Decision rules clustering using k-means algorithm with different distance measures*, in *ICSS '10 Proceedings of the 2010 International Conference*

on Systems Science: Advances in Systems Science, pp. 139–147, Academic Publishing House EXIT, 2010. [cytowanie na str. 3]

[65] Goodwin G.C., Graebe S.F., Salgado M.E., *Control system design*, Prentice-Hall, 2001. [cytowanie na str. 12]

[66] Guo F., Qian G., *Monocular 3D tracking of articulated human motion in silhouette and pose manifolds*, Journal on Image and Video Processing - Anthropocentric Video Analysis: Tools and Applications, 2008:1–18, 2008. [cytowanie na str. 7, 23, 38, 43, 66, 94]

[67] Hartley R., Zisserman A., *Multiple view geometry in computer vision*, Cambridge University Press, Cambridge, UK, 2000. [cytowanie na str. 26, 29]

[68] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer-Verlag, 2008. [cytowanie na str. 3]

[69] Heller K.A., Ghahramani Z., *Bayesian hierarchical clustering*, in *ICML '05 Proceedings of the 22nd International Conference on Machine Learning*, pp. 297–304, 2005. [cytowanie na str. 3]

[70] Holm S., *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, 6(2):65–70, 1979. [cytowanie na str. 104]

[71] Hou S., Galata A., Caillette F., Thacker N.A., *Real-time body tracking using a gaussian process latent variable model*, in *ICCV '07 Proceedings of the 11th IEEE International Conference on Computer Vision*, pp. 1–8, 2007. [cytowanie na str. 6, 23, 38, 61, 94]

[72] Howe N.R., Leventon M.E., Freeman W.T., *Bayesian reconstruction of 3D human motion from single-camera video*, in *NIPS '99 Proceedings of the Advances in Neural Information Processing Systems*, pp. 820–826, 1999. [cytowanie na str. 6, 23]

[73] Hu X.L., Schön T.B., Ljung L., *A basic convergence result for particle filtering*, IEEE Transactions on Signal Processing, 56(4):1337–1348, 2008. [cytowanie na str. 44]

[74] Hu X.L., Schön T.B., Ljung L., *A general convergence result for particle filtering*, IEEE Transactions on Signal Processing, 59(7):3424–3429, 2011. [cytowanie na str. 44]

[75] Isard M., Blake A., *CONDENSATION —- conditional density propagation for visual tracking*, International Journal of Computer Vision, 29(1):5–28, 1998. [cytowanie na str. 3, 6, 43]

[76] Jacobs R., Jordan M.I., Nowlan S.J., Hinton G.E., *Adaptive mixtures of local experts*, Neural Computation, 3:79–87, 1991. [cytowanie na str. 3, 125]

[77] Jaeggli T., Koller-Meier E., Van Gool L., *Learning generative models for multi-activity body pose estimation*, International Journal of Computer Vision, 83:121—-134, 2009. [cytowanie na str. 7, 23, 38, 66]

[78] Jepson A.D., Fleet D.J., El-Maraghi T.F., *Robust online appearance models for visual tracking*, in *CVPR '01 Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 415—422, 2001. [cytowanie na str. 81]

[79] Ji X., Liu H., *Advances in view-invariant human motion analysis: A review*, IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 40(1):13–24, 2010. [cytowanie na str. 2, 4]

[80] Jordan M.I., Ghahramani Z., Jaakkola T.S., Saul L.K., *An introduction to variational methods for graphical models*, Machine Learning, 37(2):183–233, 1999. [cytowanie na str. 3]

[81] Juszczyszyn K., Gonczarek A., Tomczak J.M., Musiał K., Budka M., *A probabilistic approach to structural change prediction in evolving social networks*, in *ASONAM '12 Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 996–1001, 2012. [cytowanie na str. 3]

[82] Kalman R.E., *A new approach to linear filtering and prediction problems*, Transactions of the American Society for Mechanical Engineering, Series D, Journal of Basic Engineering, 82:35—-45, 1960. [cytowanie na str. 37]

[83] Kanaujia A., Sminchisescu C., Metaxas D., *Semi-supervised hierarchical models for 3D human pose reconstruction*, in *CVPR '07 Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. [cytowanie na str. 4, 5]

[84] Kehl R., Van Gool L., *Markerless tracking of complex human motions from multiple views*, Computer Vision and Image Understanding, 104:190—209, 2006. [cytowanie na str. 6, 23, 61]

[85] Kirkpatrick S., Gelatt, Jr. C.D., Vecchi M.P., *Optimization by simulated annealing*, Science, 220(4598):671–680, 1983. [cytowanie na str. 50]

[86] Kjellström H., Kragić D., Black M.J., *Tracking people interacting with objects*, in *CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [cytowanie na str. 23, 43, 61, 66]

[87] Ko J., Fox D., *GP-BayesFilters: Bayesian filtering using gaussian process prediction and observation models*, Autonomous Robots, 27(1):75–90, 2009. [cytowanie na str. 50]

[88] Koller D., Friedman N., *Probabilistic Graphical Models. Principles and Techniques*, The MIT Press, London, 2009. [cytowanie na str. 36, 39]

[89] Kuipers J.B., *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace, and Virtual Reality*, Princeton University Press, Princeton, NJ, 1999. [cytowanie na str. 14, 16]

[90] Lawrence N.D., *Probabilistic non-linear principal component analysis with gaussian process latent variable models*, Journal of Machine Learning Research, 6:1783—1816, 2005. [cytowanie na str. 3, 87, 105]

[91] Lawrence N.D., Moore A.J., *Hierarchical gaussian process latent variable models*, in *ICML '07 Proceedings of the 24th international conference on Machine learning*, pp. 481–488, 2007. [cytowanie na str. 38, 93, 97, 126]

[92] Lawrence N.D., Quiñonero-Candela J., *Local distance preservation in the GP-LVM through back constraints*, in *ICML '06 Proceedings of the 23rd international conference on Machine learning*, pp. 513–520, 2006. [cytowanie na str. 91]

[93] Lawrence N.D., Seeger M., Herbrich R., *Fast sparse gaussian process methods: The informative vector machine*, in *NIPS '03 Proceedings of the Advances in Neural Information Processing Systems*, pp. 625–632, 2003. [cytowanie na str. 93]

[94] Lee C.S., Elgammal A., *Coupled visual and kinematic manifold models for tracking*, International Journal of Computer Vision, 87:118—-139, 2010. [cytowanie na str. 7, 23, 38, 41, 43, 98]

[95] Lee M.W., Cohen I., *Proposal maps driven MCMC for estimating human body pose in static images*, in *CVPR '04 Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, 2004. [cytowanie na str. 8, 82]

[96] Lee M.W., Nevatia R., *Human pose tracking in monocular sequence using multilevel structured models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1):27–38, 2009. [cytowanie na str. 8, 74, 82]

[97] Li R., Tian T., Sclaroff S., Yang M., *3D human motion tracking with a coordinated mixture of factor analyzers*, International Journal of Computer Vision, 87:170—-190, 2010. [cytowanie na str. 6, 23, 38, 41, 43, 61, 66, 82, 97, 98]

[98] Liebowitz D., Zisserman A., *Combining scene and auto-calibration constraints*, in *ICCV '99 Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999. [cytowanie na str. 27]

[99] Lindeberg T., *Scale-space theory: A basic tool for analysing structures at different scales*, Journal of Applied Statistics, 21(2):225–270, 1994. [cytowanie na str. 76]

[100] Lowe D.G., *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60(2):91–110, 2004. [cytowanie na str. 3, 76]

[101] Luong Q.T., Faugeras O.D., *Self-calibration of a moving camera from point correspondences and fundamental matrices*, International Journal of Computer Vision, 22(3):261–289, 1997. [cytowanie na str. 27]

[102] Lv F., Zhao T., Nevatia R., *Camera calibration from video of a walking human*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9):1513–1518, 2006. [cytowanie na str. 27]

[103] MacCormick J., Isard M., *Partitioned sampling, articulated objects, and interface-quality hand tracking*, in *ECCV '00 Proceedings of the 6th European Conference on Computer Vision-Part II*, pp. 3–19, 2000. [cytowanie na str. 4, 51]

[104] MacKay D.J.C., *Introduction to monte carlo methods*, in *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pp. 175–204, 1998. [cytowanie na str. 3]

[105] Mairal J., Bach F., Ponce J., Sapiro G., Zisserman A., *Supervised dictionary learning*, in *NIPS '09 Proceedings of the Advances in Neural Information Processing Systems*, pp. 1033—1040, 2009. [cytowanie na str. 3, 76]

[106] Marsland S., *Machine Learning. An Algorithmic Perspective*, Chapman & Hall. CRC Press, 2009. [cytowanie na str. 3]

[107] Memisevic R., *Kernel information embeddings*, in *ICML '06 Proceedings of the 23rd international conference on Machine learning*, pp. 633–640, 2006. [cytowanie na str. 3, 94]

[108] Memisevic R., Sigal L., Fleet D.J., *Shared kernel information embedding for discriminative inference*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4):778–790, 2012. [cytowanie na str. 4, 5, 38, 41, 66, 98]

[109] Mikić I., Trivedi M., Hunter E., Cosman P., *Human body model acquisition and tracking using voxel data*, International Journal of Computer Vision, 53(3):199–223, 2003. [cytowanie na str. 6, 61]

[110] Moeslund T.B., Hilton A., Krüger V., *A survey of advances in vision-based human motion capture and analysis*, Computer Vision and Image Understanding, 104:90—-126, 2006. [cytowanie na str. 2, 4]

[111] Mohr D., Zachmann G., *FAST: Fast adaptive silhouette area based template matching*, in *BMVC '10 Proceedings of the British Machine Vision Conference*, pp. 1–12, 2010. [cytowanie na str. 4, 68]

[112] Murphy K.P., *Machine Learning. A Probabilistic Perspective*, The MIT Press, Massachusetts Institute of Technology, 2012. [cytowanie na str. 3]

[113] Neal R.M., *Probabilistic inference using markov chain monte carlo methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, September 1993. [cytowanie na str. 43]

[114] Ning H., Tan T., Wang L., Hu W., *Kinematics-based tracking of human walking in monocular video sequences*, Image and Vision Computing, 22:429—-441, 2004. [cytowanie na str. 23, 61]

[115] Nocedal J., Wright S.J., *Numerical Optimization*, Springer-Verlag, New York, 1999. [cytowanie na str. 28, 51, 89]

[116] Okuma K., Taleghani A., de Freitas N., Little J.J., Lowe D.G., *A boosted particle filter: Multitarget detection and tracking*, in *Computer Vision - ECCV 2004. Lecture Notes in Computer Science Volume 3021*, pp. 28–39, 2004. [cytowanie na str. 50]

[117] Pearl J., *Reverend bayes on inference engines: A distributed hierarchical approach*, in *AAAI '82 Proceedings of the Second National Conference on Artificial Intelligence*, pp. 133—136, 1982. [cytowanie na str. 3]

[118] Petersen K.B., Pedersen M.S., *The Matrix Cookbook*, 2012, URL `http://matrixcookbook.com`. [cytowanie na str. 90]

[119] Peursum P., Venkatesh S., West G., *Tracking-as-recognition for articulated full-body human motion analysis*, in *CVPR '07 Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. [cytowanie na str. 43, 61]

[120] Peursum P., Venkatesh S., West G., *A study on smoothing for particle-filtered 3D human body tracking*, International Journal of Computer Vision, 87:53—-74, 2010. [cytowanie na str. 6, 41, 43, 61, 98]

[121] Piccardi M., *Background subtraction techniques: a review*, in *SMC '04 Proceedings of the 2004 IEEE Conference on Systems, Man and Cybernetics, vol.4*, pp. 3099–3104, 2004. [cytowanie na str. 3, 64]

[122] Platt J.C., *Fast training of support vector machines using sequential minimal optimization*, in *Advances in kernel methods*, pp. 185–208, MIT Press, 1999. [cytowanie na str. 78]

[123] Pollefeys M., Van Gool L., *Stratified self-calibration with the modulus constraint*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(8):707–724, 1999. [cytowanie na str. 27]

[124] Poppe R., *Vision-based human motion analysis: An overview*, Computer Vision and Image Understanding, 108:4—-18, 2007. [cytowanie na str. 2, 4]

[125] Porikli F., *Integral histogram: a fast way to extract histograms in cartesian spaces*, in *CVPR '05 Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 829–836, 2005. [cytowanie na str. 75]

[126] Psarrou A., Gong S., Walter M., *Recogntion of human gestures and behavior based on motion trajectories*, Image and Vision Computing, 20:349–358, 2002. [cytowanie na str. 4]

[127] Rasmussen C.E., Williams C.K.I., *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, 2006. [cytowanie na str. 3, 87]

[128] Rius I., Gonzàlez J., Varona J., Roca F.X., *Action-specific motion prior for efficient bayesian 3D human body tracking*, Pattern Recognition, 42:2907–2921, 2009. [cytowanie na str. 7, 23, 41, 43, 81, 98]

[129] Robert C.P., Casella G., *Monte Carlo Statistical Methods*, Springer Science+Business Media Inc., New York, 2004. [cytowanie na str. 43]

[130] Roberts T.J., McKenna S.J., Ricketts I.W., *Human tracking using 3D surface colour distributions*, Image and Vision Computing, 24(12):1332–1342, 2006. [cytowanie na str. 7]

[131] Rose C., Saboune J., Charpillet F., *Reducing particle filtering complexity for 3D motion capture using dynamic Bayesian networks*, in *AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*, pp. 1396–1401, 2008. [cytowanie na str. 43, 125]

[132] Roweis S.T., Ghahramani Z., *A unifying review of linear gaussian models*, Neural Computation, 11(2):305—-345, 1999. [cytowanie na str. 89]

[133] Roweis S.T., Saul L.K., *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290(5500):2323–2326, 2000. [cytowanie na str. 3, 94]

[134] Sapp B., Toshev A., Taskar B., *Cascaded models for articulated pose estimation*, in *ECCV'10 Proceedings of the 11th European conference on Computer vision: Part II*, pp. 406–420, 2010. [cytowanie na str. 8, 82]

[135] Schölkopf B., Smola A.J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, 2001. [cytowanie na str. 77, 78, 79]

[136] Selig J.M., *Geometric Fundamentals of Robotics*, Springer, New York, 2005. [cytowanie na str. 13, 14, 16, 23, 28]

[137] Serre T., Wolf L., Poggio T., *Object recognition with features inspired by visual cortex*, in *CVPR '05 Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–1000, 2005. [cytowanie na str. 3, 76]

[138] Shapiro L.G., Stockman G.C., *Computer Vision*, Prentice-Hall, Englewood Cliffs, NJ, 2001. [cytowanie na str. 3, 25, 26, 69]

[139] Sidenbladh H., Black M.J., Fleet D.J., *Stochastic tracking of 3D human figures using 2D image motion*, in *ECCV '00 Proceedings of the 6th European Conference on Computer Vision-Part II*, pp. 702–718, 2000. [cytowanie na str. 23, 43, 61]

[140] Sigal L., Balan A.O., Black M.J., *Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion*, International Journal of Computer Vision, 87(1):4–27, 2010. [cytowanie na str. 6, 7, 23, 41, 43, 53, 61, 65, 66, 67, 72, 82, 83, 98]

[141] Sigal L., Bhatia S., Roth S., Black M.J., Isard M., *Tracking loose-limbed people*, in *CVPR '04 Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, 2004. [cytowanie na str. 15, 61]

[142] Sigal L., Isard M., Haussecker H., Black M.J., *Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation*, International Journal of Computer Vision, 98(1):15–48, 2012. [cytowanie na str. 8, 41, 74, 82, 98]

[143] Singh V.K., Nevatia R., Huang C., *Efficient inference with multiple heterogeneous part detectors for human pose estimation*, in *ECCV'10 Proceedings of the 11th European conference on computer vision conference on Computer vision: Part III*, pp. 314–327, 2010. [cytowanie na str. 8, 82]

[144] Sminchisescu C., Triggs B., *Estimating articulated human motion with covariance scaled sampling*, International Journal of Robotics Research, 22(6):371–393, 2003. [cytowanie na str. 6, 7, 61]

[145] Stauffer C., Grimson W.E.L., *Adaptive background mixture models for real-time tracking*, in *CVPR '99 Proceedings of the 1999 IEEE Conference on Computer Vision and Pattern Recognition*, 1999. [cytowanie na str. 3, 64]

[146] Stefanov N., Galata A., Hubbold R., *A real-time hand tracker using variable-length markov models of behaviour*, Computer Vision and Image Understanding, 108:98—-115, 2007. [cytowanie na str. 4]

[147] Sudderth E., Ihler A., Freeman W., Willsky A., *Nonparametric belief propagation*, in *CVPR '03 Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 605—-612, 2003. [cytowanie na str. 3]

[148] Sun M., Kohli P., Shotton J., *Conditional regression forests for human pose estimation*, in *CVPR '12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [cytowanie na str. 82]

[149] Sun M., Telaprolu M., Lee H., Savarese S., *An efficient branch-and-bound algorithm for optimal human pose estimation*, in *CVPR '12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [cytowanie na str. 8, 82]

[150] Sundaresan A., Chellappa R., *Model-driven segmentation of articulating humans in laplacian eigenspace*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(10):1771–1785, 2008. [cytowanie na str. 61]

[151] Sundaresan A., Chellappa R., *Multicamera tracking of articulated human motion using shape and motion cues*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(9):2114–2126, 2009. [cytowanie na str. 23, 61]

[152] Świątek J., *Wybrane zagadnienia identyfikacji statycznych systemów złożonych*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2009. [cytowanie na str. 32]

[153] Taylor G.W., Sigal L., Fleet D.J., Hinton G.E., *Dynamical binary latent variable models for 3D human pose tracking*, in *CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [cytowanie na str. 6, 23, 35, 40, 41, 43, 61, 66, 97, 98]

[154] Tenenbaum J.B., de Silva V., Langford J.C., *A global geometric framework for nonlinear dimensionality reduction*, Science, 290(5500):2319–2323, 2000. [cytowanie na str. 3, 94]

[155] Tian T., Li R., Sclaroff S., *Tracking human body pose on a learned smooth space*, Technical Report 2005-029, Boston University Computer Science Department, August 2005. [cytowanie na str. 6, 38, 43, 94]

[156] Tipping M.E., *Sparse bayesian learning and the relevance vector machine*, Journal of Machine Learning Research, 1:211–244, 2001. [cytowanie na str. 79]

[157] Titsias M.K., Lawrence N.D., *Bayesian gaussian process latent variable model*, in *AISTATS'10 Proceedings of the 13th International Conference on Articial Intelligence and Statistics*, 2010. [cytowanie na str. 93]

[158] Tomczak J.M., Gonczarek A., *Decision rules extraction from data stream in the presence of changing context for diabetes treatment*, Knowledge and Information Systems, 34(3):521–546, 2013. [cytowanie na str. 79]

[159] Tsai R.Y., *A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses*, IEEE Journal on Robotics and Automation, 3(4):323–344, 1987. [cytowanie na str. 27]

[160] Tuzel O., Porikli F., Meer P., *Region covariance: a fast descriptor for detection and classification*, in *ECCV'06 Proceedings of the 9th European conference on Computer Vision - Volume Part II*, pp. 589–600, 2006. [cytowanie na str. 75]

[161] Urtasun R., Fleet D.J., Fua P., *3D people tracking with gaussian process dynamical models*, in *CVPR '06 Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, 2006. [cytowanie na str. 6, 23, 38, 81, 94]

[162] Urtasun R., Fleet D.J., Fua P., *Temporal motion models for monocular and multiview 3D human body tracking*, Computer Vision and Image Understanding, 104:157—177, 2006. [cytowanie na str. 6]

[163] Urtasun R., Fleet D.J., Geiger A., Popović J., Darrell T.J., Lawrence N.D., *Topologically-constrained latent variable models*, in *ICML '08 Proceedings of the 25th international conference on Machine learning*, pp. 1080–1087, 2008. [cytowanie na str. 38, 93]

[164] Urtasun R., Fleet D.J., Hertzmann A., Fua P., *Priors for people tracking from small training sets*, in *ICCV '05 Proceedings of the Tenth IEEE International Conference on Computer Vision*, pp. 403–410, 2005. [cytowanie na str. 38, 94]

[165] Vapnik V.N., *Statistical Learning Theory*, John Wiley & Sons, Inc., 1998. [cytowanie na str. 32, 77, 78]

[166] Vapnik V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 2000. [cytowanie na str. 77, 78]

[167] Viola P., Jones M.J., *Robust real-time face detection*, International Journal of Computer Vision, 57(2):137—-154, 2004. [cytowanie na str. 3, 74, 75]

[168] Wainwright M.J., Jordan M.I., *Graphical models, exponential families, and variational inference*, Foundations and Trends in Machine Learning, 1(1–2):1–305, 2008. [cytowanie na str. 36, 39]

[169] Wang J., Fleet D.J., Hertzmann A., *Gaussian process dynamical models for human motion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2):283–298, 2008. [cytowanie na str. 38, 40, 93, 94]

[170] Weinberger K.Q., Saul L.K., *Unsupervised learning of image manifolds by semidefinite programming*, International Journal of Computer Vision, 70(1):77–90, 2006. [cytowanie na str. 3, 94]

[171] Werghi N., *Segmentation and modeling of full human body shape from 3-D scan data: A survey*, IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 37(6):1122–1136, 2007. [cytowanie na str. 3]

[172] Yang Y., Ramanan D., *Articulated pose estimation with flexible mixtures-of-parts*, in *CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [cytowanie na str. 8, 82, 126]

[173] Yilmaza A., Shahb M., *A differential geometric approach to representing the human actions*, Computer Vision and Image Understanding, 109(3):335—-351, 2008. [cytowanie na str. 4]

[174] Zhang Z., *Flexible camera calibration by viewing a plane from unknown orientations*, in *ICCV '99 Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999. [cytowanie na str. 29]

[175] Zhang Z., *A flexible new technique for camera calibration*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330–1334, 2000. [cytowanie na str. 27, 28]

[176] Zhang Z., *Camera calibration with one-dimensional objects*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(7):892–899, 2004. [cytowanie na str. 27]

[177] Zhao X., Liu Y., *Generative tracking of 3D human motion by hierarchical annealed genetic algorithm*, Pattern Recognition, 41:2470—2483, 2008. [cytowanie na str. 6, 23]

Spis symboli i skrótów

Symbol/skrót	Opis
MOCAP	System Motion Capture
MAP	Estymator maksymalnego prawdopodobieństwa a posteriori
(x, y, z)	Lokalny ortogonalny układy współrzędnych
$x - y - z$	Kolejność wykonywania obrotów wokół osi
θ_x, θ_y, θ_z	Kąty Eulera dla poszczególnych osi
$\mathbf{n}$	Jednostkowy wektor z twierdzenia Eulera
n_x, n_y, n_z	Składowe wektora $\mathbf{n}$
φ	Kąt obrotu z twierdzenia Eulera
$\mathbf{q}$	Jednostkowy kwaternion
q_w, q_x, q_y, q_z	Składowe jednostkowego kwaternionu
$\|\| \cdot \|\|$	Norma euklidesowa
$\overline{\mathbf{q}}$	Zredukowana postać kwaternionu zdefiniowana przez (2.6)
$\mathbf{v}$	Punkt w przestrzeni $\mathbb{R}^3$
v_x, v_y, v_z	Składowe punktu $\mathbf{v}$
$\mathbf{R}$	Macierz rotacji w przestrzeni $\mathbb{R}^3$
$\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z$	Kolumny macierzy rotacji
$\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$	Wersory rozpinające standardowy układ współrzędnych
$\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z$	Macierze rotacji wokół ustalonych osi układu (x, y, z)
r_{ij}	Elementy macierzy rotacji $\mathbf{R}$
$\breve{\mathbf{q}}$	Wektor złożony z trzech ostatnich składowych kwaternionu zdefiniowany przez (2.17)
$\mathbf{Q}$	Macierz zdefiniowana przez (2.18)
$\mathbf{I}_{n\times n}$	Macierz jednostkowa o wymiarach $n \times n$

Symbol/skrót	Opis
$\mathcal{V}$	Zbiór punktów elementu sztywnego wyrażony w lokalnym układzie współrzędnych
(X, Y, Z)	Globalny układ współrzędnych
$\mathbf{u}$	Wektor przesunięcia w przestrzeni $\mathbb{R}^3$
$\overline{\mathbf{v}}$	Punkt $\mathbf{v}$ wyrażony w globalnym układzie współrzędnych
$\overline{\mathcal{V}}$	Element sztywny wyrażony w globalnym układzie współrzędnych
$\mathbf{R}_i$	Macierz rotacji dla i-tego elementu sztywnego pomiędzy lokalnym i globalnym układem współrzędnych
$\mathbf{u}_i$	Wektor przesunięcia dla i-tego elementu sztywnego pomiędzy lokalnym i globalnym układem współrzędnych
$\mathbf{v}_i$	Punkt $\mathbf{v}$ wyrażony w lokalnym układzie i-tego elementu
$\mathbf{R}_{i,j}$	Macierz rotacji z lokalnego układu i-tego elementu do lokalnego układu j-tego elementu
$\mathbf{u}_{i,j}$	Położenie początku lokalnego układu i-tego elementu w lokalnym układzie j-tego elementu
$\mathrm{pa}(i)$	Indeks rodzica i-tego elementu
$\mathcal{V}_i$	i-ty element sztywny w drzewie kinematycznym
$\overline{\mathcal{V}}_i$	i-ty element sztywny wyrażony w globalnym układzie współrzędnych
h_i	Długość i-tej kończyny
$\boldsymbol{\theta}_i$	Kąty Eulera odpowiadające macierzy rotacji $\mathbf{R}_{i,\mathrm{pa}(i)}$
$\theta_{i,x}, \theta_{i,y}, \theta_{i,z}$	Składowe wektora $\boldsymbol{\theta}_i$
$\check{\mathbf{x}}$	Zredukowany wektor stanu zdefiniowany przez (2.35) lub (2.37)
$\boldsymbol{\theta}_0$	Kąty Eulera odpowiadające macierzy rotacji $\mathbf{R}_0$
$\theta_{0,x}, \theta_{0,y}, \theta_{0,z}$	Składowe wektora $\boldsymbol{\theta}_0$
$\mathbf{x}$	Wektor stanu zdefiniowany przez (2.36) lub (2.38)
$\overline{\mathbf{q}}_i$	Zredukowana postać kwaternionu odpowiadająca obrotowi $\boldsymbol{\theta}_i$
I	Obraz wejściowy z pojedynczej kamery
I_{ij}^c	Piksel o współrzędnych (i, j) i kolorze c
$\mathcal{I}$	Zbiór obrazów ze wszystkich dostępnych kamer

Symbol/skrót	Opis
$\mathbf{v}^{\mathrm{I}}$	Punkt $\mathbf{v}$ wyrażony układzie współrzędnych kamery odpowiadającej za obraz I
$\tilde{\mathbf{v}}^{\mathrm{I}}$	Projekcja punktu $\mathbf{v}^{\mathrm{I}}$ na oraz I
$\tilde{v}_x^{\mathrm{I}}, \tilde{v}_y^{\mathrm{I}}$	Współrzędne punktu $\tilde{\mathbf{v}}^{\mathrm{I}}$ na obrazie I
$\mathbf{A}$	Macierz parametrów wewnętrznych kamery
s	Dodatni parametr skali w zadaniu kalibracji kamery
$\alpha_c, \beta_c, \gamma_c, a_c, b_c$	Parametry wewnętrzne kamery
$\tilde{\mathbf{v}}$	Rozszerzona reprezentacja punktu w globalnym układzie współrzędnych
$\mathbf{R}_{\mathrm{I}}$	Macierz rotacji pomiędzy globalnym układem współrzędnych i lokalnym układem kamery
$\mathbf{u}_{\mathrm{I}}$	Położenie początku globalnego układu w lokalnym układzie kamery
$\tilde{\mathbf{v}}_n$	Punkt wyróżniony na szablonie kalibracyjnym w globalnym układzie współrzędnych i w rozszerzonej postaci
$\tilde{\mathbf{v}}_n^{\mathrm{I}}$	Punkt wyróżniony na szablonie kalibracyjnym po zrzutowaniu na obraz I
$\mathcal{P}(\mathbf{v})$	Projekcja perspektywiczna punktu $\mathbf{v}$ określona przez zależność (2.44)
$\mathcal{P}_{\mathrm{I}}(\overline{\mathbf{v}})$	Projekcja punktu $\overline{\mathbf{v}}$ na obraz I zdefiniowana przez zależność (2.45)
$\mathcal{P}_{\mathrm{I}}(\mathcal{V})$	Projekcja elementu sztywnego na obraz I zdefiniowana przez zależność (2.46)
$p(\cdot)$	Gęstość rozkładu prawdopodobieństwa
$\hat{\mathbf{x}}$	Estymata wektora stanu $\mathbf{x}$
$R[\cdot]$	Funkcjonał ryzyka
$L(\cdot,\cdot)$	Funkcja straty
$\delta(\cdot)$	Delta Diraca
$p(\cdot\|\cdot)$	Gęstość warunkowego rozkładu prawdopodobieństwa
$\mathbb{E}[\cdot\|\cdot]$	Warunkowa wartość oczekiwana
$\mathbf{x}_t$	Wektor stanu w chwili t
$\mathbf{x}_{t_1:t_2}$	Sekwencja wektorów stanu od chwili t_1 do t_2
$\mathcal{I}_t$	Zbiór obrazów ze wszystkich dostępnych kamer w chwili t

Symbol/skrót	Opis
$\mathcal{I}_{t_1:t_2}$	Sekwencja zbiorów obrazów od chwili t_1 do t_2
$\mathbf{x}_1 \perp \mathbf{x}_2 \mid \mathbf{x}_3$	Niezależność zmiennych losowych $\mathbf{x}_1$ i $\mathbf{x}_2$ pod warunkiem zmiennej losowej $\mathbf{x}_3$
$\hat{\mathbf{x}}_{t_1:t_2}$	Sekwencja estymat wektora stanu od chwili t_1 do t_2
$\hat{\mathbf{x}}_t$	Estymata wektora stanu w chwili t
$\mathbf{z}$	Punkt w układzie współrzędnych związanym z niskowymiarową rozmaitością
$\mathbf{z}_t$	Punkt w układzie współrzędnych związanym z niskowymiarową rozmaitością odpowiadający wektorowi stanu w chwili t
$\mathbf{z}_{t_1:t_2}$	Sekwencja punktów $\mathbf{z}_t$ od chwili t_1 do t_2
$\mathcal{W}$	Zbiór wyróżnionych punktów testowych
$\mathbf{w}_i(\mathbf{x})$	Położenie i-tego punktu testowego w globalnym układzie współrzędnych w zależności od wektora stanu $\mathbf{x}$
$\mathrm{err}(\cdot)$	Błąd śledzenia ruchu zdefiniowany przy pomocy zależność (3.28)
$\mathbf{x}^{(n)}$	Pojedyncza realizacja wygenerowana z rozkładu na wektor stanu
$\hat{p}(\cdot)$	Aproksymacja rozkładu $p(\cdot)$
$\mathbb{E}[\cdot]$	Wartość oczekiwana
$\tilde{\pi}(\mathbf{x})$	Wartość funkcji wiarygodności wyliczona dla wektora stanu $\mathbf{x}$
$\hat{p}(\cdot\|\cdot)$	Aproksymacja rozkładu warunkowego $p(\cdot\|\cdot)$
$\pi(\mathbf{x})$	Unormowana postać $\tilde{\pi}(\mathbf{x})$ zgodnie z zależnością (4.7)
$\mathcal{X}^{\pi}$	Zbiór cząsteczek zdefiniowany przez (4.8)
$\mathbf{x}_t^{(n)}$	Pojedyncza realizacja wygenerowana z rozkładu na wektor stanu w chwili t
$\mathcal{X}_t^{\pi}$	Zbiór cząsteczek w chwili t zdefiniowany przez (4.15)
$\overline{\mathbf{x}}_t^{(n)}$	Pojedyncza realizacja wygenerowana z dyskretnej aproksymacji rozkładu na wektor stanu w chwili t
$\mathcal{X}_t$	Próba z rozkładu a priori na wektor stanu w chwili t
$\overline{\mathcal{X}}_t$	Próba z dyskretnej aproksymacji rozkładu a posteriori na wektor stanu w chwili t
β_i	Parametr wyżarzania w i-tej warstwie

Symbol/skrót	Opis
$\mathbf{x}_{t,l}^{(n)}$	Pojedyncza realizacja wygenerowana z rozkładu na wektor stanu w chwili t i warstwie wyżarzania l
$\pi_i(\mathbf{x})$	Waga dla wektora stanu $\mathbf{x}$ w i-tej warstwie wyżarzania określona zależnością (4.19)
$ESS(\cdot)$	Kryterium określające efektywną wielkość próby zdefiniowane zależnością (4.20)
$\alpha_p(\cdot)$	Odsetek efektywnych cząsteczek zdefiniowany przez (4.21)
α_p	Pożądany odsetek efektywnych cząsteczek
$\mathcal{X}_{t,l}$	Próba z rozkładu a priori na wektor stanu w chwili t i l-tej warstwie wyżarzania
$\overline{\mathcal{X}}_{t,l}$	Próba z dyskretnej aproksymacji rozkładu a posteriori na wektor stanu w chwili t i l-tej warstwie wyżarzania
$q(\cdot\|\cdot)$	Gęstość pomocniczego warunkowego rozkładu prawdopodobieństwa
Z	Czynnik normujący zadany przez zależność (4.25)
$\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)$	Współczynniki wagowe zdefiniowane przy pomocy zależności (4.27)
$\mathbf{z}_t^{(n)}$	Pojedyncza realizacja wygenerowana z rozkładu na wektor $\mathbf{z}$ w chwili t
$\hat{q}(\cdot\|\cdot)$	Aproksymacja pomocniczego rozkładu warunkowego $q(\cdot\|\cdot)$
$\omega(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)$	Unormowane parametry wagowe zdefiniowane przez (4.35)
$\mathcal{Z}_t$	Próba z rozkładu na wektor $\mathbf{z}$ w chwili t
h, ρ_1, ρ_2	Poszczególne wymiary ściętego stożka
$\mathcal{B}$	Model ciała zdefiniowany jako zbiór elementów $\mathcal{V}$
D^{I}	Mapa głębi dla obrazu I
$\mathrm{D}_{ij}^{\mathrm{I}}$	Wartość piksela (i, j) na mapie głębi
$\mathcal{J}^{\mathrm{I}}(\mathcal{V})$	Zbiór indeksów pikseli (i, j), na których widoczny jest element $\mathcal{V}$ po zrzutowaniu na obraz I zdefiniowany przez (5.3)
$\mathbf{c}_i$	Środek elementu sztywne $\mathcal{V}_i$ w globalnym układzie współrzędnych
d_i^{I}	Odległość $\mathbf{c}_i$ od kamery z obrazem I
$\mathcal{J}^{\mathrm{I}}(\mathbf{x})$	Zbiór pikseli, na których widoczny jest model ciała zdefiniowany przez (5.4)

Symbol/skrót	Opis
p_0	Wartość progowa w metodzie oddzielania tła
$\mathcal{N}(\cdot\|\cdot,\cdot)$	Gęstość rozkładu normalnego
$\mathcal{I}_{BG}$	Zbiór obserwacji tła dla ustalonej kamery
S^{I}	Binarna sylwetka – wynik procedury oddzielania tła
$\mathrm{S}^{\mathrm{I}}_{ij}$	Wartość piksela (i,j) na binarnej sylwetce
$\tilde{p}(\cdot\|\cdot)$	Gęstość warunkowego rozkładu prawdopodobieństwa w nieunormowanej postaci
$\mathcal{E}^{\mathrm{I}}(\mathcal{V})$	Zbiór indeksów pikseli (i,j), na które zrzutowane zostały punkty rozłożone wzdłuż krawędzi elementu $\mathcal{V}$
$\mathcal{E}^{\mathrm{I}}(\mathbf{x})$	Zbiór pikseli, na których widoczne są punkty rozłożone wzdłuż krawędzi modelu ciała zdefiniowany przez (5.13)
$\upsilon^{\mathrm{I}}_{ij}$	Wskaźnik stwierdzający czy punkt z ustalonego zbioru jest widoczny na pikselu (i,j) po zrzutowaniu na obraz I
$\mathrm{F}^x, \mathrm{F}^y$	Filtry gradientowe
$\mathrm{G}^{\mathrm{I},x}, \mathrm{G}^{\mathrm{I},y}$	Obraz I po przefiltrowaniu odpowiednio F^x i F^y
G^{I}	Binarna mapa krawędzi zdefiniowana przez (5.21)
η	Wartość progowa w wyrażeniu (5.21)
F^g	Filtr gaussowski
σ_g	Parametr filtra gaussowskiego
E^{I}	Mapa krawędzi zdefiniowana przez (5.23)
e_0	Stała występująca w modelu wiarygodności opartym na krawędziach
$\mathcal{M}^{\mathrm{I}}(\mathcal{V})$	Zbiór trójek (i,j,k), gdzie (i,j) oznacza położenie unikatowego punktu k na obrazie I dla elementu $\mathcal{V}$
$\mathcal{M}^{\mathrm{I}}(\mathbf{x})$	Zbiór pikseli, na których widoczne są wszystkie unikalne punkty wyróżnione na ciele zdefiniowany przez (5.25)
$\boldsymbol{\phi}$	Deskryptor fragmentu obrazu
ϕ^i	i-ta składowa deskryptora
H^i	Filtry w postaci falek Haara
ϕ^m_{ij}	Składowa deskryptora dla punktu (i,j) zdefiniowana przez (5.26)
$\boldsymbol{\phi}_{ij}$	Deskryptor punktu (i,j)
II	Obraz całkowy zdefiniowany przez (5.27)

Symbol/skrót	Opis
h_{ca}	Wartość filtru Haara na spójnym obszarze
SVM	Klasyfikator Support Vector Machine
$\mathcal{D}_\phi$	Ciąg treningowy dla modelu wyglądu zdefiniowany przez (5.29)
y^i	Wartość ze zbioru $\{-1, 1\}$
a_i	i-ty parametr klasyfikatora SVM zdefiniowanego przez (5.32)
C	parametr regularyzacji w zadaniu uczenia (5.30)
$k(\cdot, \cdot)$	Funkcja jądra zdefiniowana przez (5.31)
σ_ϕ	Parametr funkcji jądra (5.31)
$\mathbf{a}$	Wektor parametrów klasyfikatora SVM
SV	Zbiór indeksów oznaczających wektory wspierające
$s(\cdot)$	Siła reakcji klasyfikatora SVM określona przez zależność (5.32)
b	Parametr przesunięcia określony przez (5.33)
s_0	Wartość progowa dla klasyfikatora SVM
$s_k(\cdot)$	Model wyglądu dla k-tego punktu
v^{I}_{ijk}	Wskaźnik stwierdzający czy unikatowy punkt k jest widoczny na pikselu (i, j) po zrzutowaniu na obraz I
$\sigma(\cdot)$	Funkcja sigmoidalna zdefiniowana przez (5.34)
σ_0	Stała w modelu wiarygodności opartym na lokalnych deskryptorach (5.36)
λ_i	Wagi dla składowych modeli wiarygodności
$\boldsymbol{\varepsilon}_t$	Szum gaussowski o niezależnych składowych
$\mathrm{diag}(\cdot)$	Macierz diagonalna o diagonali zadanie przez wektor
$\boldsymbol{\sigma}^2$	Wektor parametrów podstawowego modelu dynamiki zdefiniowanego przez (6.3)
$\mathcal{D}_x$	Ciąg treningowy zawierający wektory stanu otrzymane z danych z systemu MOCAP
x^i_t	i-ta składowa wektora stanu $\mathbf{x}_t$
$\hat{F}_i(x)$	Dystrybuanta empiryczna zdefiniowana przez (6.5)
$\mathbb{1}(\cdot)$	Indykator
α	Wartość z przedziału $[0, 1]$ zadająca poziom istotności
$\kappa_i(\alpha)$	Kwantyl rzędu α zdefiniowany przez (6.7)

Symbol/skrót	Opis
$\mathbf{u}_{0,t}$	Globalne położenie drzewa kinematycznego w chwili t
$\boldsymbol{\theta}_{0,t}$	Globalna rotacja drzewa kinematycznego w chwili t
$\check{\mathbf{x}}_t$	Zredukowany wektor stanu w chwili t
$\boldsymbol{\sigma}^2_{\mathbf{u}}, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}$	Wektory parametrów dla modeli dynamiki dla globalnego położenia i rotacji
$\mathbf{X}$	Macierz zawierająca w wierszach zredukowane wektory stanu ze zbioru treningowe $\mathcal{D}_x$
$\mathbf{Z}$	Macierz zawierająca w wierszach niskowymiarowe reprezentacje dla zredukowanych wektorów stanu ze zbioru treningowego $\mathcal{D}_x$
GPLVM	Gaussian Process Latent Variable Model
ε	Niezależny szum gaussowski z modelu GPLVM
$\mathbf{f}(\cdot)$	Wektor funkcji zadający odwzorowanie pomiędzy niskowymiarową reprezentacją i zredukowanym wektorem stanu
f_i	Składowa wektora $\mathbf{f}$
$\mathcal{GP}(\cdot\|\cdot,\cdot)$	Proces Gaussa
$k_z(\cdot, cdot)$	Funkcja kowariancji zdefiniowana przez (6.16)
σ_z^2	Wariancja szumu w modelu GPLVM
β, β_0, γ_z	Parametry funkcji kowariancji
$\mathbf{f}_i$	Wektor wartości funkcji f_i dla wszystkich przykładów z macierzy $\mathbf{Z}$
$\mathbf{F}$	Macierz zdefiniowana przez zależność (6.17)
k_{nm}	Wartość funkcji kowariancji dla para przykładów zdefiniowana przez (6.19)
$\mathbf{K}$	Macierz kowariancji złożona z elementów k_{nm}
$\check{x}_t^i$	Składowe zredukowanego wektora stanu $\check{\mathbf{x}}_t$
$\mathbf{X}_{:,i}$	i-ta kolumna macierzy $\mathbf{X}$
$\|\cdot\|$	Wyznacznik macierzy
$\mathrm{tr}(\cdot)$	Ślad macierzy
$\overline{\mathbf{K}}$	Macierz kowariancji $\mathbf{K}$ poprawiona o wariancję szumu σ_z^2
$\|\cdot\|_F$	norma Frobeniusa
$L(\cdot)$	Funkcja celu w procesie uczenia modelu GPLVM zdefiniowana przez (6.25)

Symbol/skrót	Opis
$\overline{k}_z(\cdot,\cdot)$	Funkcja kowariancji zdefiniowana przez (6.28)
δ_{nm}	Delta Kroneckera
BC-GPLVM	Back-Constrained Gaussian Process Latent Variable Model
$\mathbf{g}(\cdot)$	Wektor funkcji zadający odwzorowanie pomiędzy zredukowanym wektorem stanu i niskowymiarową reprezentacją
g_i	Składowa wektora $\mathbf{g}$
c_{ti}, b_i	Parametry modelu dla odwzorowania g_i zadanego zależnością (6.32)
$k_x(\cdot,\cdot)$	Funkcja jądra zadana przez (6.33)
γ_x	Parametr funkcji jądra $k_x(\cdot,\cdot)$
$\overline{\mathbf{k}}$	Wektor zawierający wartości funkcji kowariancji $\overline{k}_z(\cdot,\cdot)$ wyliczone poprzez porównanie przykładu $\mathbf{z}$ z przykładami w macierzy $\mathbf{Z}$
$\boldsymbol{\mu}_p, \sigma_p^2$	Wektor średniej i wariancja rozkładu predykcyjnego wyrażonego za pomocą (6.37)
$\boldsymbol{\varepsilon}_t^{x\to z}$	Szum gaussowski o niezależnych składowych w modelu (6.40)
$\boldsymbol{\sigma}^2_{x\to z}$	Parametry modelu (6.42)
$\boldsymbol{\sigma}^2_{x\to x}$	Parametry modelu (6.44)
$\boldsymbol{\varepsilon}_t^{z\to x}$	Szum gaussowski o niezależnych składowych w modelu (6.45)
$\boldsymbol{\sigma}^2_{z\to x}$	Parametry modelu (6.48)
MPF	Filtr cząsteczkowy uwzględniający strukturę rozmaitości
SIR	Zwykły filtr cząsteczkowy (Sampling Importance Resampling)
APF	Wyżarzany filtr cząsteczkowy
H_i	Hipoteza statystyczna
m_i	Mediana błędów i-tej metody
S	Model oparty na sylwetkach
S+E	Połączenie modeli opartego na sylwetkach i opartego na krawędziach
BS	Model oparty na dwustronnych sylwetkach
LD	Model oparty na lokalnych deskryptorach
BS+LD	Połączenie modeli opartego na dwustronnych sylwetkach i opartego na lokalnych deskryptorach

Wytłuszczone symbole odnoszą się do wektorów i macierzy.

Spis rysunków

1.1 Idea systemu do bezznacznikowego odtwarzania konfiguracji ciała człowieka. 2

2.1 Reprezentacje obrotu w przestrzeni trójwymiarowej. (a) Kąty Eulera. (b) Kwaterniony. 14
2.2 Transformacja elementu sztywnego z lokalnego do globalnego układu współrzędnych. 20
2.3 Obiekty przegubowo połączone. (a) Pojedyczny staw. (b) Drzewo kinematyczne dla człowieka. 22
2.4 Projekcja perspektywiczna. (a) Projekcja obiektu na ekran. (b) Identyczny obraz dla różnych obiektów. 29

3.1 Probabilistyczny model grafowy dla ukrytego modelu Markowa. 35
3.2 Układ współrzędnych na niskowymiarowej rozmaitości. 39
3.3 Probabilistyczny model grafowy uwzględniający strukturę niskowymiarowej rozmaitości. 39

4.1 Schemat działania filtra cząsteczkowego. 48
4.2 Schemat działania filtra cząsteczkowego uwzględniającego strukturę niskowymiarowej rozmaitości. 58

5.1 Modelowanie ciała człowieka. (a) Element sztywny w postaci ściętego stożka. (b) Model ciała. (c) Mapa głębi. 61
5.2 Model wiarygodności oparty na sylwetkach. (a) Wejściowy obraz I. (b) Binarna sylwetka S^{I}. (c) Porównanie sylwetki z obrazu wejściowego z sylwetką ze zrzutowanego modelu ciała. 65

5.3 Model wiarygodności oparty na krawędziach. (a) Wejściowy obraz I. (b) Mapa krawędzi E^{I}. (c) Porównanie mapy krawędzi z obrazu wejściowego z krawędziami ze zrzutowanego modelu ciała. 71
5.4 Filtry oparte na falkach Haara. (a) Rodzaje filtrów. (b) Przykładowe położenia filtra w otoczeniu punktu. 74
5.5 Model wiarygodności oparty na lokalnych deskryptorach. 81

7.1 Zbiory punktów w niskowymiarowym układzie uzyskane na podstawie ciągów treningowych dla rozważanych sekwencji ruchu. Osie oznaczają odpowiednio współrzędne z^1 i z^2. 102
7.2 Punkty testowe służące do oceny poprawności algorytmu śledzącego. 103
7.3 Rozkład błędów śledzenia (3.28) dla rozważanych sekwencji. 106
7.4 Przebieg średniego błędu w kolejnych klatkach śledzenia dla rozważanych sekwencji. 109
7.5 Przykładowy przebieg śledzenia dla wyżarzanego filtra cząsteczkowego (APF). Sekwencja S3-Jog. Kamera nr 2. 110
7.6 Przykładowy przebieg śledzenia dla zwykłego filtra cząsteczkowego (SIR). Sekwencja S3-Jog. Kamera nr 2. 110
7.7 Przykładowy przebieg śledzenia dla filtra cząsteczkowego uwzględniającego strukturę rozmaitości (MPF). Sekwencja S3-Jog. Kamera nr 2. 111
7.8 Rozkład błędów śledzenia (3.28) dla rozważanych sekwencji i testowanych modeli wiarygodności. 115
7.9 Przebieg średniego błędu w kolejnych klatkach śledzenia dla rozważanych sekwencji i wybranych modeli wiarygodności. 119
7.10 Przykładowy przebieg śledzenia dla modelu opartego na sylwetkach (S). Sekwencja S1-Walk. Kamera nr 2. 120
7.11 Przykładowy przebieg śledzenia dla modelu opartego na sylwetkach i krawędziach (S+E). Sekwencja S1-Walk. Kamera nr 2. 120
7.12 Przykładowy przebieg śledzenia dla modelu opartego na dwustronnych sylwetkach (BS). Sekwencja S1-Walk. Kamera nr 2. 121
7.13 Przykładowy przebieg śledzenia dla modelu opartego na lokalnych deskryptorach (LD). Sekwencja S1-Walk. Kamera nr 2. 121

7.14 Przykładowy przebieg śledzenia dla modelu opartego na dwustronnych sylwetkach i lokalnych deskryptorach (BS+LD). Sekwencja S1-Walk. Kamera nr 2. 122

Spis tabel

7.1 Wyodrębnione sekwencje ruchu na potrzeby badań empirycznych 99
7.2 Błąd śledzenia ruchu [mm] określony zależnością (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji. 107
7.3 Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm]. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach.108
7.4 Różnice średnich błędów pomiędzy metodami z literatury i referencyjną metodą MPF dla rozważanych sekwencji testowych. Rezultaty testu statystycznego dla poziomu istotności $\alpha = 0.05$. 111
7.5 Błąd śledzenia ruchu [mm] dla sekwencji z chodzeniem zadany przy pomocy zależności (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji. 114
7.6 Błąd śledzenia ruchu [mm] dla sekwencji z bieganiem zadany przy pomocy zależności (3.28). Pogrubiono najlepsze wyniki dla każdej sekwencji. 116
7.7 Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm] dla sekwencji z chodzeniem. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach. 117
7.8 Średni błąd śledzenia ruchu (3.28) dla wyróżnionych fragmentów ciała [mm] dla sekwencji z bieganiem. Pogrubiono najlepsze wyniki dla każdej części ciała w wszystkich sekwencjach. 118
7.9 Różnice średnich błędów pomiędzy modelami znanymi z literatury i referencyjnym modelem BS+LD dla rozważanych sekwencji testowych. Rezultaty testu statystycznego dla poziomu istotności $\alpha = 0.05$. 122

Skorowidz

A*-search, 8
AdaBoost, 79
algorytm BFGS, 89
analiza głównych składowych, 6, 90
Automatic Relevance Determination, 79

Back-Constrained Gaussian Process Latent Variable Model, 91, 95

Canny Edge Detector, 69
CONDENSATION, 43
corse-to-fine, 8
Covarianced Scaled Sampling, 6

detekcja krawędzi, 69
drzewo kinematyczne, 5–8, 23, 24, 41, 60

estymacja pozy, 4, 8, 10, 12, 25, 31, 33, 34, 37, 44, 66, 82, 98
Expectation-Maximization, 8

falki Haara, 74, 76, 126
filtr cząsteczkowy, 6, 9, 10, 38, 41, 43, 44, 47, 49, 50, 52–54, 57–59, 67, 98, 100, 101, 105, 113, 124, 125
filtr cząsteczkowy Rao-Blackwella, 49
filtr Kalmana, 37, 50
filtrowanie, 9, 37, 38, 40, 41, 43, 44, 47, 54, 56, 58, 124
funkcja sigmoidalna, 80

Gaussian Process Dynamical Model, 93
Gaussian Process Latent Variable Model, 6, 87, 92–94, 97, 108, 126
Generative Topographic Mapping, 94
Geometric blur, 76
Graph-based Rules Inducer, 79

Hierarchical Hidden Markov Model, 6
Histogram of oriented gradients, 76
HMAX, 76
HumanEva, 65, 99, 100, 103
Hyperfeatures, 76

Informative Vector Machine, 93
Interior-Point, 78
Isomap, 93

kąty Eulera, 13, 14, 17–19, 24, 25
kalibracja kamer, 26, 29, 100
Kernel Information Embedding, 94
klątwa wymiarowości, 25, 50
klasteryzacja, 3, 6, 97
kwaternion, 14–16, 18–20, 24, 25, 85, 92, 101

Laplacian Eigenmaps, 94
liniowy model gaussowski, 89
Locally Linear Embedding, 94

mapa głębi, 61, 68
Markov Chain Monte Carlo, 8
markowskie pole losowe, 7
maszyna Boltzmanna, 76, 97
Maximum Variance Unfolding, 94
metoda gradientów sprzężonych, 89
metoda Monte Carlo, 43
metoda podziału i ograniczeń, 8

metody spektralne, 93
mieszanina analiz czynnikowych, 6
mieszanina ekspertów, 5, 125
mieszanina rozkładów Gaussa, 6
model ciała, 7, 26, 60, 61, 63, 66–73, 78, 79, 81, 103, 105, 110, 116, 118, 121
model dynamiki, 10, 37, 40, 41, 48, 49, 83, 85, 86, 94, 124–126
model dyskryminacyjny, 4, 5, 33, 66
model generujący, 4–8, 33, 125
model oparty na częściach, 5, 7, 8, 82, 125
model wiarygodności, 6–11, 33, 38, 41, 48, 60, 66–68, 71, 72, 76, 79, 80, 82, 98, 101, 105, 112, 113, 123–126
model wyglądu, 26, 73, 74, 78, 79, 112, 114, 117, 122, 125

Nonparametric Belief Propagation, 8

obraz całkowy, 75
oddzielanie tła, 3, 64–67, 101, 105
ograniczenia wsteczne, 91, 92, 94
ograniczona maszyna Boltzmanna, 6
Out-Of-Sample, 95

podejście bottom-up, 7, 82
podejście top-down, 6, 7, 82
ponowne próbkowanie, 47, 49, 57
próbkowanie znaczące, 45
probabilistyczny model grafowy, 35, 39
proces Gaussa, 5, 87, 88, 93
programowanie dynamiczne, 7
programowanie kwadratowe, 78
przepływ optyczny, 7
przetwarzanie obrazów, 69

Random Forests, 79
redukcja wymiarów, 3, 6, 59, 95, 97
regresja jądrowa, 95, 97
regresja liniowa, 4
regresja logistyczna, 79
regularyzacja, 4, 59, 77, 79, 89, 90, 112
Relevance Vector Machine, 79
rozkład predykcyjny, 92, 95
rozmaitość, 6, 7, 9–11, 38–40, 54, 58, 59, 66, 85–87, 91–98, 100, 101, 105, 107, 124–126
rozpoznawanie akcji, 4
rozszerzony filtr Kalmana, 6

Scale-invariant feature transform, 76
sekwencyjne Monte Carlo, 43
sekwencyjne próbkowanie znaczące, 49
Sequential Minimal Optimization, 78
Shared Gaussian Process Latent Variable Model, 5, 93
Shared Kernel Information Embedding, 5
skompresowane zrozumienie, 64
statystyczna teoria uczenia, 77
Stochastic Meta Descent, 6
struktura obrazkowa, 7, 8
Support Vector Machine, 4, 77–79, 114
sylwetka, 4, 7, 60, 63, 65–69, 71, 72, 101, 105, 107, 113, 122, 123, 125
symulowane wyżarzanie, 6, 50
system dynamiczny, 12
system MOCAP, 1, 16, 41, 83, 86, 96, 98–100, 112

śledzenie ruchu człowieka, 1–4, 8–12, 18, 25, 31, 37, 38, 40, 42–44, 46–50, 52, 66, 67, 82, 83, 86, 93, 98, 103, 105, 112, 113, 116, 124, 125

twierdzenie Bayesa, 33
twierdzenie Eulera, 14, 18

uczenie bez nadzoru, 6
uczenie maszynowe, 3, 10, 35, 87
uczenie słowników, 76
uczenie z częściowym nadzorem, 5
ukryty model Markowa, 31, 35, 97

Variable Length Markov Model, 6, 126

Wandering-Stable-Lost, 81
wektor stanu, 12, 13, 24, 25, 31–42, 46, 47, 50, 53, 63, 66, 80, 82–85, 87, 91, 92, 97, 98, 101, 124, 125
wektor wspierający, 78, 79
widzenie komputerowe, 3, 10, 26, 64, 74, 75
wyżarzany filtr cząsteczkowy, 6, 43, 50, 52–54, 101
wymiar Vapnika-Chervonenkisa, 77