ZESZYTY NAUKOWE POLITECHNIKI RZESZOWSKIEJ 294, Elektrotechnika 35
RUTJEE, z. 35 (1/2016), styczeń-marzec 2016, s. 51-67

Dawid WARCHOŁ[1]

PORÓWNANIE I ŁĄCZENIE CECH DESKRYPTORÓW CHMUR PUNKTÓW W ROZPOZNAWANIU STATYCZNYCH UKŁADÓW DŁONI

W pracy omówiono metodę rozpoznawania statycznych układów dłoni przy użyciu trzech deskryptorów chmur punktów: Viewpoint Feature Histogram (VFH), Global Radius-Based Surface Descriptor (GRSD) oraz Ensemble of Shape Functions (ESF). Deskryptory te opisują chmury punktów wykorzystując siatkę prostopadłościennych, wolumetrycznych elementów (ang. voxel grid), wektory normalne wyznaczone dla powierzchni chmury, rozkłady odległości punktów od ich sąsiadów oraz promienie sfer wpisanych we fragmenty powierzchni. Przeprowadzono testy walidacji krzyżowej uzyskując w ten sposób zestawienie wyników skuteczności klasyfikacji dla pojedynczych cech każdego deskryptora, łączonych cech tego samego i różnych deskryptorów. Testy przeprowadzono na zbiorze danych zawierającym 1000 map głębi: 10 różnych układów dłoni wykonanych 10 razy przez 10 osób. Przed przystąpieniem do procesu ekstrakcji cech każda chmura punktów jest wstępnie przetwarzana, włączając w to: segmentację (w celu oddzielenia dłoni od pozostałych fragmentów chmury), rotację względem środka dłoni oraz najdłuższego z wysuniętych palców (w celu uniezależnienia metody od obrotów dłoni wokół osi prostopadłej do obiektywu kamery) oraz redukcję punktów (w celu przyspieszenia obliczeń). Zestawienie wyników testów uzupełniono o dodatkową informację – rozmiar wektora cech wykorzystanego przy klasyfikacji, co pozwala odnaleźć kombinację cech będącą punktem kompromisu pomiędzy skutecznością klasyfikacji a ilością wymiarów danych.

Słowa kluczowe: deskryptory chmur punktów, Histogram Cech Zależnych od Punktu Widzenia, Globalny Deskryptor Powierzchni Bazujący na Promieniach Sfer, Zestaw Funkcji Kształtu

1. Wstęp

W artykule tym omówiono podejście do rozpoznawania postur dłoni (nazywanych też statycznymi gestami) w oparciu o trójwymiarowe dane głębi w formie chmur punktów.

[1] Dawid Warchoł, Politechnika Rzeszowska im. Ignacego Łukasiewicza, Wydział Elektrotechniki i Informatyki, ul. W. Pola 2, 35-021 Rzeszów, tel. 796 795 080, email: dawwar@prz.edu.pl.

Najbardziej popularnym i niedrogim urządzeniem do pozyskiwania danych głębi jest sensor Microsoft Kinect™, wprowadzony w listopadzie 2010 roku jako kontroler gier komputerowych. Wartymi uwagi są również kamery typu time-of-flight, wcześniej uważane za drogie, specjalistyczne urządzenia, które obecnie stają się coraz bardziej przystępne dla rynku masowego. Rosnąca popularność tych urządzeń spowodowała zainteresowanie badaczy tematyką rozpoznawania gestów wykonywanych dłonią z wykorzystaniem kamer głębi. Najczęściej jednak dane 3D są używane jedynie do segmentacji lub jako informacja pomocnicza zawarta w wektorach cech klasyfikowanych obiektów (np. [11], [12], [19], [21]). Kombinacja danych kolorowych oraz głębi została zastosowana przez [20] w celu klasyfikacji postur dłoni wykorzystując transformację Average Neighborhood Margin Maximization aproksymowaną Haarletami, czyli cechami opartymi na falkach Haara. Odmienne podejście do rozpoznawania układów dłoni bazując na danych głębi zostało przedstawione w [6]. Informacja pozyskana z kamery Kinect™ została użyta w celu wygenerowania szkieletów przez algorytm Mean Shift Local Mode Finding. Następnie zastosowano metodę dopasowania szkieletów oraz losowe lasy decyzyjne aby dokonać klasyfikacji pikseli głębi na fragmenty dłoni. Znaczący wpływ danych głębi na cechy obiektów wykorzystywane w klasyfikacji można zaobserwować w pracy [3]. Deskryptor opracowany przez autorów zawiera następujące zestawy cech: (i) odległości czubków palców od środka dłoni, (ii) odległości czubków palców od płaszczyzny aproksymującej powierzchnię dłoni, (iii) krzywizna konturów dłoni oraz (iv) kształt środkowej części dłoni (bez palców).

W niniejszym artykule proponujemy wykorzystanie informacji 3D w postaci deskryptorów chmur do klasyfikacji postur dłoni. Rozpatrujemy trzy deskryptory: (i) Viewpoint Feature Histogram (VFH), (ii) Global Radius-Based Surface Descriptor (GRSD), (iii) Ensemble of Shape Functions (ESF). Zbadana zostanie skuteczności klasyfikacji dla pojedynczych cech każdego deskryptora, łączonych cech tego samego i różnych deskryptorów. Rozmiar wektora cech wykorzystanego przy klasyfikacji w każdym z testowanych przypadków pozwoli odnaleźć kombinację cech będącą punktem kompromisu pomiędzy skutecznością klasyfikacji a ilością wymiarów danych.

Deskryptor VFH był przedmiotem naszych dotychczasowych badań, natomiast pozostałe 2 deskryptory nie były wcześniej wykorzystywane do rozpoznawania gestów wykonywanych dłonią. W [4] zaproponowano modyfikację sposobu liczenia VFH w celu rozpoznawania gestów dynamicznych, którą następnie wykorzystano do rozpoznawania statycznych postur dłoni [5]. Polegała ona na dzieleniu obserwowanej sceny na mniejsze prostopadłościenne komórki i wyznaczaniu cech deskryptora dla każdej z nich (niezależnie od pozostałych). Metoda ta zwiększyła dystynktywność VFH, szczególnie w przypadku obiektów o subtelnych różnicach kształtu, co zostało potwierdzone poprzez eksperymenty, które wykazały znacznie większą skuteczność rozpoznawania w przypadku dzie-

lonej sceny. Z tego powodu w niniejszej pracy wykorzystano metodę dzielenia sceny do obliczania cech każdego z trzech deskryptorów.

2. Deskryptory chmur punktów

W rozdziale tym przedstawiono teoretyczny opis wszystkich trzech deskryptorów chmur punktów wykorzystywanych w niniejszej pracy. W przeprowadzonych eksperymentach (opisanych w rozdziale 3) wykorzystywaliśmy bibliotekę PCL do przetwarzania chmur punktów oraz obliczania wartości deskryptorów.

2.1. Viewpoint Feature Histogram

Viewpoint Feature Histogram (VFH, pol. Histogram Cech Zależnych od Punktu Widzenia) po raz pierwszy został przedstawiony w [15]. Jest to globalny deskryptor chmury punktów, czyli struktury reprezentującej wielowymiarowy zbiór punktów w układzie współrzędnych [16]. Horyzontalna oś *x* układu jest skierowana w lewo, wertykalna oś *y* jest skierowana ku górze, natomiast oś *z* pokrywa się z osią optyczną kamery i jest zwrócona w stronę obserwowanego obiektu. VFH składa się z dwóch komponentów: (i) kształtu powierzchni, który opisuje geometryczne właściwości obiektu oraz (ii) kierunku patrzenia. Deskryptor ten jest w stanie wykryć subtelne zmiany geometrii obiektów nawet w przypadku powierzchni nieteksturowanych, co zostało dowiedzione eksperymentalnie [15].

Komponent kształtu powierzchni składa się z cech θ, $\cos(\alpha)$, $\cos(\Phi)$ i d mierzonych między środkiem ciężkości chmury p_c i każdym punktem p_i do niej należącym (patrz rys. 1). n_c jest wektorem z punktem przyłożenia p_c, którego współrzędne są równe średniej wszystkich wektorów normalnych do powierzchni, natomiast n_i reprezentuje normalne oszacowane w punkcie p_i. Kąty θ i α można opisać jako kątami odchylenia (ang. yaw) oraz pochylenia (ang. pan) pomiędzy dwoma wektorami, natomiast d określa odległość euklidesową pomiędzy p_i i p_c. Wektory i kąty przedstawione na rysunku 1 zdefiniowane są następująco:

$$u = n_c \tag{1}$$

$$v = \frac{p_i - p_c}{d} \times u \tag{2}$$

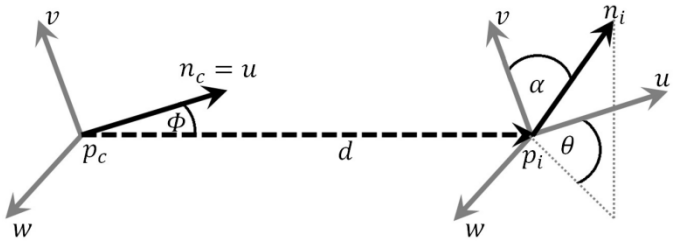
$$w = u \times v \tag{3}$$

$$\cos(\alpha) = v \cdot n_i \tag{4}$$

$$\cos(\Phi) = u \cdot \frac{p_i - p_c}{d} \tag{5}$$

$$\theta = arctg\left(\frac{w \cdot n_i}{u \cdot n_i}\right) \tag{6}$$

gdzie kropka oznacza iloczyn skalarny, natomiast krzyżyk – iloczyn wektorowy.



Rys. 1. Cechy komponentu kształtu powierzchni VFH
Fig. 1. Features of the surface shape component of the VFH

Komponent kierunku patrzenia zawiera histogram kątów pomiędzy linią będącą kierunkiem patrzenia (pokrywającą się z osią *z*) a każdym wektorem normalnym. Metoda obliczania VFH ma jeden parametr nn_{vfh} oznaczający liczbę punktów należących do lokalnego sąsiedztwa, który jest wykorzystywany do oszacowania wektorów normalnych. Histogramy VFH zawierają po 45 przedziałów dla każdej cechy komponentu kształtu powierzchni oraz 128 dla komponentu kierunku patrzenia (w sumie 308 przedziałów). Bardziej szczegółowy opis obliczania VFH jest przedstawiony w [18], [14] (deskryptory PFH oraz FPFH) oraz [15].

2.2. Global Radius-Based Surface Descriptor

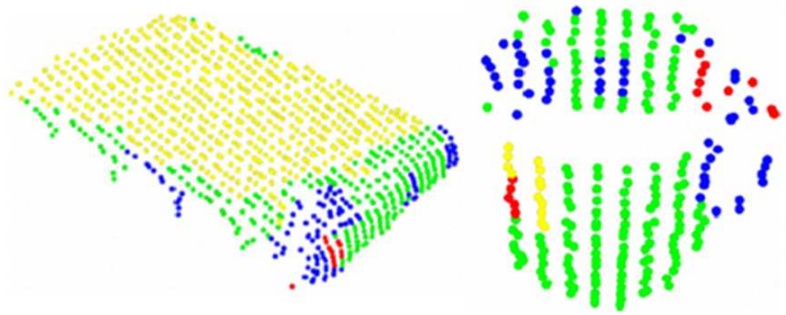
Deskryptor Global Radius-based Surface Descriptor (GRSD, pol. Globalny Deskryptor Powierzchni Bazujący na Promieniach Sfer) po raz pierwszy pojawił się w [10]. Jest on globalną wersją deskryptora Radius-Based Surface Descriptor (RSD, pol. Deskryptor Powierzchni Bazujący na Promieniach Sfer) powstałą z połączenia lokalnego RSD, deskryptora Global Fast Point Feature Histogram (GFPFH, pol. Globalny Szybki Histogram Cech Punktów) [17] oraz metody 2D Speeded Up Robust Features (SURF) [1]. GRSD opisuje radialne relacje punktów z ich sąsiedztwem. Aby zrozumieć jego działanie, należy najpierw przedstawić lokalny deskryptor RSD, który został opisany w pracach [8], [9].

RSD jest tworzony w następujący sposób. Dla każdej pary punktu i jego sąsiada algorytm oblicza odległości między nimi oraz różnicę pomiędzy ich wek-

torami normalnymi. Następnie wyznaczana jest sfera, na której powierzchni leżą oba punkty. Do określenia położenia i promienia sfery wykorzystywana jest również informacja o wektorach normalnych punktów: wektory te muszą być prostopadłe do powierzchni sfery. Pominięcie informacji o normalnych spowodowałoby istnienie nieskończonej ilość sfer dla każdej pary punktów. Na koniec dla danego punktu wybrane zostają dwie sfery: największa i najmniejsza, których promienie tworzą deskryptor tego punktu.

Można zaobserwować, że w przypadku, gdy oba punkty leżą na zakrzywionej ścianie walca, promień wygenerowanej sfery będzie mniej więcej równy promieniowi walca. Jeśli natomiast punkty leżą na jednej płaszczyźnie, promień sfery będzie nieskończenie długi. Algorytm przyjmuje 2 parametry: rnp_{grsd} – promień lokalnego sąsiedztwa, który jest wykorzystywany do oszacowania wektorów normalnych oraz rnn_{grsd} – promień, w obrębie którego wybierani są sąsiedzi punktów.

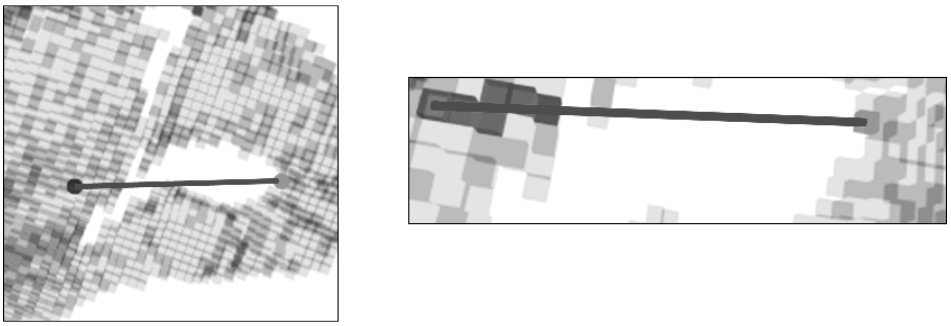
Aby wyznaczyć globalną wersję deskryptora RSD, czyli GRSD, należy najpierw przeprowadzić kategoryzację powierzchni. Obliczone w każdym punkcie deskryptory RSD są wprowadzane jako wejście algorytmu Conditional Random Field (CRF), który przypisuje punktowi jedną z pięciu etykiet kategorii: płaszczyzna, walec, ostra krawędź (lub szum), obrzeże, sfera. Mamy więc na tym etapie chmurę z każdym punktem sklasyfikowanym w zależności od typu obiektu lub pewnego regionu, do którego punkt należy (rys. 2).



Rys. 2. Kategoryzacja powierzchni przeprowadzona w każdym punkcie dwóch obiektów; różne kolory reprezentują następujące kategorie: płaszczyzna - żółty, walec - zielony, ostra krawędź lub szum - czerwony, obrzeże (np. granica przejścia pomiędzy powierzchniami) - niebieski, sfera (nie została sklasyfikowana w przedstawionych chmurach punktów); na podstawie [10]

Fig. 2. Surface categorization performed for every point in the two objects; different colors represent the following categories: plane - yellow, cylinder - green, sharp edge or noise - red, rim (boundary, transition between surfaces) - blue, sphere (not classified in the presented point clouds); based on [10]

Kolejnym krokiem jest utworzenie tzw. drzewa ósemkowego (ang. octree), czyli hierarchicznej struktury przechowywania wokseli (trójwymiarowego odpowiednika pikseli o sześciennym kształcie) ułatwiającej ich przeszukiwanie, kompresję oraz próbkowanie chmury punktów. Dla każdego utworzonego woksela wyznaczamy kategorię powierzchni, którą z największym prawdopodobieństwem reprezentuje. Jest to kategoria, do której należących punktów jest najwięcej wewnątrz danego woksela. Następnie, dla każdej pary wokseli w całym octree, przeprowadza się linię łączącą ich środki (rys. 3). Kolejnym krokiem jest znalezienie wokseli, przez które przechodzi wyznaczona linia i sprawdzenie, czy znajduje się w nich przynajmniej jeden punkt. Na tej podstawie tworzony jest wykres, który każdemu wokselowi przypisuje numer jego kategorii lub 0, jeśli woksel jest pusty.



Rys. 3. Obliczanie GRSD na siatce wokseli; rysunek przedstawia przykładową linię łączącą 2 woksele utworzone na podstawie chmury punktów; na podstawie [17]

Fig. 3. The calculation of the GRSD descriptor on a voxel grid; the figure presents an example line connecting 2 woksels created based on a point cloud; based on [17]

Otrzymane wykresy mają zmienną długość osi *x*. Na ich podstawie tworzony jest jeden globalny histogram o stałej długości. Każdy przedział tego histogramu reprezentuje inną kombinację tranzycji (przejścia) pomiędzy wokselami różnych kategorii (włączając kategorię 0), przy czym pary <*kategoria*$_a$, *kategoria*$_b$> oraz <*kategoria*$_b$, *kategoria*$_a$> są równoważne. Przedziały histogramu są wypełniane liczbami poszczególnych tranzycji występujących we wszystkich wykresach. Rozmiar histogramu jest równy liczbie 2-elementowych kombinacji z powtórzeniami ze zbioru *nc*-elementowego, gdzie *nc* jest liczbą kategorii powierzchni, włączając kategorię 0 (wokseli pustych): $(nc + 2) \cdot (nc + 1)/2$.

Dla *nc* = 6 liczba ta jest równa 21, a więc GRSD jest opisywany przez histogram składający się z 21 przedziałów.

2.3. Ensemble of Shape Functions

Deskryptor Ensemble of Shape Functions został (ESF, pol. Zestaw Funkcji Kształtu) przedstawiony w [21]. Jest on kombinacją trzech różnych funkcji kształtu opisujących następujące właściwości chmury punktów: odległości i kąty pomiędzy punktami oraz obszary utworzone przez poszczególne punkty. ESF jest unikalny, ponieważ nie wymaga obliczania normalnych do powierzchni, ani żadnego przetwarzania wstępnego z wyjątkiem utworzenia siatki wokseli chmury, która aproksymuje rzeczywistą powierzchnię. Deskryptor składa się z 10 histogramów, z których każdy zawiera 64 przedziały. Algorytm tworzenia histogramów wykonuje się zadaną liczbę esf_n razy (domyślnie wartość ta równa jest 20000). W każdej iteracji losowane są 3 punkty p_1, p_2, p_3, dla których obliczane są cechy: *D2*, *D2-ratio*, *D3*, *A3*.

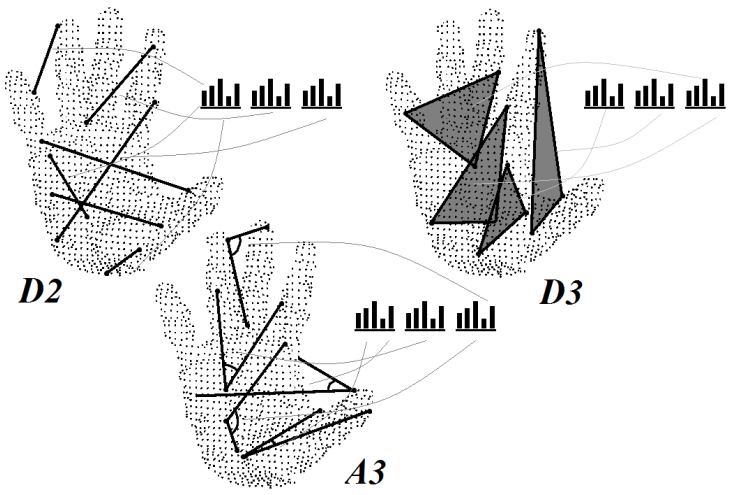
D2 jest cechą składającą się z trzech histogramów. Do jej wyznaczenia obliczane są odległości pomiędzy punktami p_1, p_2, p_3. Następnie dla każdej pary algorytm sprawdza, czy linia łącząca punkty leży: a) całkowicie na powierzchni chmury, b) całkowicie poza nią (z wyjątkiem punktów początkowego i końcowego), czy c) częściowo na powierzchni, częściowo poza nią. W zależności od tego, wartość będąca odległością punktów umieszczana jest w jednym z trzech histogramów: IN, OUT lub MIXED.

D2-ratio jest opisana przez pojedynczy histogram, którego wartości określają stosunek między długościami fragmentów linii leżących na powierzchni i poza nią. Jedynie linie zakwalifikowane do histogramu MIXED są brane pod uwagę.

D3 obejmuje 3 histogramy. Wyznaczenie tej cechy polega na obliczeniu pierwiastka kwadratowego powierzchni trójkąta, którego wierzchołki znajdują się w punktach p_1, p_2, p_3. Wartości te są umieszczane w trzech histogramach, podobnie jak w przypadku cechy *D2*, w zależności od tego, czy boki trójkąta leżą w całości na powierzchni chmury, w całości poza nią, czy częściowo na i częściowo poza powierzchnią.

A3 składa się z trzech histogramów, których wygenerowanie polega na wyznaczeniu kątów pomiędzy liniami utworzonymi z połączeń punktów p_1, p_2, p_3. Wartości te są umieszczane w trzech histogramach, analogicznie do cechy *D3*, czyli w zależności od pokrycia powierzchni trójkątów utworzonych przez trójki punktów z powierzchnią chmury.

Rys. 4 ilustruje metodę wyznaczania poszczególnych cech. Autorzy ESF twierdzą, że deskryptor radzi sobie z różnicami w charakterystyce kamer, z których pobierane są chmury punktów, umożliwiając stworzenie solidnej metody klasyfikacji obiektów 3D.



Rys. 4. Cechy ESF;
Fig. 4. Features of the ESF;

3. System rozpoznawania układów dłoni

Opracowany system rozpoznawania układów dłoni składa się z następujących elementów: (i) segmentacja dłoni, (ii) rotacja dłoni, (iii) konwersja mapy głębi na chmurę punktów, (iv) próbkowanie (ang. down-sampling), (v) wyznaczenie bryły brzegowej i jej podział na komórki (vi) ekstrakcja oraz normalizacja cech, (vii) klasyfikacja.

Segmentacja przeprowadzana jest w celu oddzielenia dłoni od pozostałych fragmentów chmury punktów tak, aby cechy deskryptora wyznaczane były jedynie dla obiektu zainteresowania (dłoni). W naszym systemie rozpoznawania zastosowano algorytm segmentacji opracowany w [3] z pewnymi modyfikacjami, których szczegóły wykraczają poza tematykę niniejszego artykułu.

Rotacja chmury wokół osi *z* przeprowadzana jest w taki sposób, aby wektor o początku znajdującym się w środku dłoni oraz końcu w jej najbardziej wysuniętym punkcie (najczęściej jest to najdłuższy z wystawionych palców) miał kierunek zgodny z osią *y*. Pozwala to uniezależnić metodę rozpoznawania dłoni od jej obrotów wokół osi *z*. Segmentację i rotacje przeprowadza się na mapach głębi, czyli dwuwymiarowych tablicach, których elementy określają głębię w danym punkcie. Jeśli więc dane zapisane są w postaci chmury punktów, należy je wczytać jako mapa głębi przed wykonaniem punktu (i) oraz (ii).

Po przeprowadzeniu tych operacji mapa głębi konwertowana jest na chmurę punktów. Współrzędne chmury punktów PC_i^x, PC_i^y i PC_i^z są ustawiane w zależności od wartości mapy głębi DA_i^z zgodnie z parametrami kamery (z której pozyskano dane) oraz równaniami rzutowania perspektywicznego:

$$PC_i^x = \frac{(DA_i^z + fl) * \left(\frac{DA_{width}}{2} - DA_i^x - 1\right) * ps^x}{fl} \tag{7}$$

$$PC_i^y = \frac{(DA_i^z + fl) * \left(\frac{DA_{height}}{2} - DA_i^y - 1\right) * ps^y}{fl} \tag{8}$$

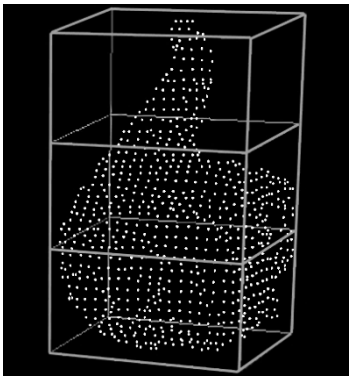
$$PC_i^z = DA_i^z \tag{9}$$

gdzie: DA_{width}– liczba kolumn mapy głębi,
DA_{width}– liczba wierszy mapy głębi,
fl – długość ogniskowej kamery,
ps^x i ps^y – wymiary pikseli, odpowiednio wysokość i szerokość.

W przypadku naszych eksperymentów, dane zostały pobrane z kamery Kinect™, której parametry pobrano z [22] i ustawiono następująco: fl = 4.73 mm, $ps^x = ps^y$ = 0.0078 mm.

Kolejnym etapem systemu rozpoznawania układów dłoni jest redukcja chmury punktów poprzez próbkowanie. Polega ono na zmniejszeniu liczby punktów wchodzących w skład nadmiarowo gęstej chmury i stosuje się go, aby przyspieszyć obliczenia związane z wyznaczaniem deskryptorów. W tym celu, w miejscach w których znajdują się punkty tworzy się siatkę prostopadłościennych wokseli, które są reprezentowane przez pojedynczy punkt zlokalizowany w ich środku ciężkości. Punkty te tworzą nową, zredukowaną chmurę. Parametrem metody są wymiary wokseli $V_x \times V_y \times V_z$. W naszych eksperymentach zdecydowaliśmy się użyć wokseli sześciennych o wymiarach: $V_x = V_y = V_z = 0.0045$ m, ponieważ nie ma przesłanek, żeby poszczególne wymiary nie były sobie równe. Drogą eksperymentalną stwierdzono, że taki rozmiar wokseli jest wystarczający do obliczenia reprezentatywnych histogramów rozpatrywanych deskryptorów.

Aby wyznaczyć cechy deskryptorów, należy utworzyć bryłę brzegową, czyli najmniejszy prostopadłościan kompletnie otaczający chmurę punktów. Bryła brzegowa wpasowuje się w chmurę nie tworząc przestrzeni pomiędzy najbardziej wystającymi punktami a ścianą. Aby zwiększyć dystynktywność deskryptorów, bryła brzegowa jest dzielona na prostopadłościenne komórki i cechy deskryptora liczone są dla każdej z nich niezależnie. Nasze eksperymenty przeprowadzone w [5] pokazały, że dzielenie bryły brzegowej na 3 horyzontalne komórki (patrz rys. 5) daje najlepsze rezultaty w rozpoznawaniu gestów wykonywanych dłonią, np. statycznych postur dłoni (sprawdzono również 3 wertykalne komórki, 9 komórek oraz brak podziału). W związku z tym w naszych badaniach wykorzystano ten właśnie podział.



Rys. 5. Zredukowana chmura punktów przedstawiająca dłoń z wystawionym kciukiem oraz jej bryła brzegowa podzieloną na trzy horyzontalne komórki

Fig. 5. Downsampled point cloud corresponding to an arm with extended thumb and its bounding box divided to three horizontal cells

Pojedyncze cechy deskryptora VFH składają się z jednego histogramu o rozmiarze 45 przedziałów, natomiast cechy ESF – z trzech histogramów po 64 przedziały. Aby uniknąć problemu związanego z przekleństwem wymiarowości, wynikającego ze zbyt dużej liczby wymiarów danych [2], histogramy reprezentowane są przez ich średnią i odchylenie standardowe. Następnie wartości te są normalizowane do zakresu [0-1] względem maksimum i minimum znajdującym się w zbiorze treningowym i w takiej postaci umieszczane są w wektorze cech. Deskryptor GRSD zawiera tylko jeden histogram o wielkości 21 przedziałów. W tym przypadku testy przeprowadzono zarówno dla całych histogramów, gdzie pojedyncze wartości przedziałów stanowiły cechy w wektorze, oraz dla ich reprezentacji poprzez średnią i odchylenie standardowe.

Rozmiar wektora cech, oznaczonego jako FV_{size}, wylicza się następująco:

$$FV_{size} = nc \cdot \sum_{i=1}^{nf} nh_i \cdot nr_i \tag{10}$$

gdzie: nc – liczba komórek, na które dzielona jest bryła brzegowa,
nf – liczba cech deskryptorów,
nh_i – liczba histogramów, z których składa się *i*-ta cecha,
nr_i – liczba reprezentantów histogramów *i*-tej cechy (wartość 2 w przypadku reprezentacji histogramów jako średnia i odchylenie standardowe).

Przykładowo, gdy wektor cech składa się z jednej cechy VFH, jednej ESF i jednej GRSD (dla przypadku reprezentowania histogramu przez każdy przedział, tj. bez liczenia średniej i odchylenia standardowego), rozmiar wektora cech jest równy $3 \cdot ((1 \cdot 2) + (3 \cdot 2) + (1 \cdot 21)) = 87$. Warto zwrócić uwagę,

że duży rozmiar wektora cech, oprócz wspomnianego zjawiska przekleństwa wymiarowości, jest również niekorzystny pod względem prędkości uczenia klasyfikatora oraz samej klasyfikacji.

Ostatnim etapem systemu rozpoznawania układów dłoni jest klasyfikacja. Do tego celu wykorzystaliśmy klasyfikator *k*-najbliższych sąsiadów z odległością euklidesową i parametrem $k = 1$ [7].

4. Wyniki eksperymentów

Eksperymenty przeprowadzone w ramach niniejszej pracy polegały na testowaniu pojedynczych cech każdego deskryptora oraz łączonych cech jednego i różnych deskryptorów w rozpoznawaniu statycznych układów dłoni. Wykorzystany został zbiór danych przedstawiony w [13]. Zawiera on 10 układów dłoni pokazywanych 10 razy przez 10 osób, co daje łącznie 1000 plików w formie map głębi. Wszystkie gesty statyczne dłoni zawarte w używanym zbiorze danych zostały przedstawione na rys. 5. Eksperymenty polegały na wykonywaniu w każdym przypadku 5-krotnej walidacji krzyżowej uzyskując w ten sposób procentowy wynik skuteczności klasyfikacji. Dla każdego testu walidacyjnego dzielono zbiór danych na 5 rozłącznych podzbiorów, z których 1 pełnił rolę zbioru testowego, a pozostałe 4 stanowiły zbiór treningowy. Podział zbioru danych przeprowadzany był w taki sposób, aby w zbiorze testowym nie znajdowały się gesty pokazane przez osoby występujące w zbiorze treningowym. Dzięki temu system rozpoznawania gestów może być testowany pod kątem właściwości uogólniających i niezależności od osoby.

Tabela 1 przedstawia wyniki eksperymentów przeprowadzonych dla pojedynczych cech każdego z deskryptorów uzupełnione o wielkość wektora w cech dla każdego przypadku. Testy walidacji krzyżowej zostały wykonane jednokrotnie dla cech deskryptorów FVH oraz GRSD, natomiast w przypadku ESF przeprowadzono 30 identycznych testów, a następnie obliczono wartość średnią wyników. Powodem tego działania jest niedeterministyczny charakter metody liczenia histogramów ESF, który wynika z wielokrotnego losowania punktów p_1, p_2, p_3 i skutkuje niewielkimi różnicami w histogramach wygenerowanych dla tej samej chmury punktów. Oznacza to, że wyniki dla każdego testu przeprowadzonego w identycznych warunkach mogą się różnić. Aby wykazać, że różnice te są niewielkie i akceptowalne, w każdym przypadku oprócz średniej obliczono również odchylenie standardowe z wyników 30 testów walidacji.

Rys. 6. Układy dłoni (wraz z etykietami) zawarte w zbiorze danych użytym w celu ewaluacji metody rozpoznawania

Fig. 6. Postures (with labels) included in the dataset used for the evaluation of the recognition method

Tabela 1. Wyniki testów 5-krotnej walidacji krzyżowej; pojedyncze cechy deskryptorów; skuteczność klasyfikacji oraz odchylenie standardowe wyników podano w procentach

Table 1. Results of 5-fold cross-validation tests; single descriptor features; recognition rates and standard deviations of results are given in percentages

Cecha deskryptora	*Skuteczność klasyfikacji*	*Odch. std. testów (tylko ESF)*	*Rozmiar wektora cech*
VFH(θ)	50.8	-	6
VFH(α)	40.6	-	6
VFH(Φ)	62.2	-	6
VFH(d)	76.1	-	6
ESF(D2)	86.64	0.43	18
ESF(D2-ratio)	54.61	0.59	6
ESF(D3)	81.71	0.51	18
ESF(A3)	88.46	0.5	18
GRSD(całe histogramy)	56	-	63
GRSD(średnia i odch. std. Histogramów)	44.8	-	6

Najlepszą w rozpoznawaniu gestów statycznych cechą VFH okazała się *d* - rozkład odległości punktów. W przypadku ESF najskuteczniejsza okazała się cecha *A3*, analizująca kąty pomiędzy liniami utworzonymi z trójek punktów. Uzyskany w tym przypadku wynik poprawności klasyfikacji – 88.46% jest najlepszym rezultatem spośród wszystkich analizowanych cech deskryptorów. Histogramy GRSD reprezentowane za pomocą średniej i odchylenia standardowe-

go poskutkowały stosunkowo niską skutecznością klasyfikacji, podczas gdy wynik uzyskany poprzez zawarcie całych histogramów w wektorze cech jest o 11.2% większy. W tym przypadku jednak wektor cech jest bardzo liczny (63 cechy), szczególnie w porównaniu do pozostałych cech deskryptorów.

Wyniki eksperymentów przeprowadzonych przy użyciu łączonych cech tych samych oraz różnych deskryptorów zostały przedstawione w Tabeli 2. Ze względu na dużą liczbę analizowanych cech, przedstawiono jedynie wybrane połączenia cech (np. połączenia cech, które okazały się najlepsze w poprzednich eksperymentach). W przypadku połączeń zawierających cechy ESF przeprowadzono 30 testów 5-krotnej walidacji krzyżowej w sposób analogiczny do eksperymentów z pojedynczymi cechami deskryptorów.

Tabela 2. Wyniki testów 5-krotnej walidacji krzyżowej; łączone cechy deskryptorów; skuteczność klasyfikacji oraz odchylenie standardowe wyników podano w procentach

Table 2. Results of 5-fold cross-validation tests; combined descriptor features; recognition rates and standard deviations of results are given in percentages

Cechy deskryptorów	*Skuteczność klasyfikacji*	*Odch. std. testów (tylko ESF)*	*Rozmiar wektora cech*
VFH(Φ)+VFH(d)	90.9	-	12
VFH(θ)+VFH(Φ)+VFH(d)	89.4	-	18
VFH(θ)+VFH(α)+VFH(Φ)+VFH(d)	92.3	-	24
VFH(θ)+VFH(α)+VFH(Φ)	75.1	-	18
ESF(D2)+ESF(D2-ratio)	87.59	0.31	24
ESF(A3)+ESF(D2)	91.8	0.48	36
ESF(A3)+ESF(D2-ratio)	90.76	0.49	24
ESF(A3)+ESF(D2)+ESF(D2-ratio)	92.27	0.37	42
ESF(A3)+ESF(D3)+ESF(D2)+ ESF(D2-ratio)	91.95	0.47	60
ESF(A3)+VFH(d)	93.61	0.31	24
ESF(A3)+VFH(Φ)+VFH(d)	94.79	0.44	30
ESF(A3)+ VFH(θ)+VFH(α)+VFH(Φ)+VFH(d)	95.28	0.33	42
GRSD(całe histogramy)+ESF(A3)	81.98	0.47	81
GRSD(całe histogramy)+VFH(d)	83.7	-	69

W przypadku VFH najskuteczniejszym okazało się połączenie wszystkich cech komponentu kształtu powierzchni ($\theta+d+\Phi+\alpha$). Należy jednak zauważyć, że wybierając dwie najlepsze cechy z poprzednich eksperymentów ($\Phi+d$) uzyskujemy poprawność klasyfikacji mniejszą jedynie o 1.4%, natomiast rozmiar wektora cech zmniejsza się dwukrotnie (z 24 na 12). Deskryptor ESF

w całości (wszystkie cechy) uzyskał skuteczność klasyfikacji zbliżoną do komponentu kształtu powierzchni VFH, jednak przy 2.5 razy większym wektorze cech. Połączenie cech różnych deskryptorów ESF($A3$)+VFH(d) poskutkowało poprawnością klasyfikacji 93.61, natomiast wektor cech ma ten sam rozmiar, który uzyskujemy w przypadku połączenia wszystkich cech komponentu kształtu powierzchni VFH. Nieco większe wektory cech oraz lepsze skuteczności klasyfikacji uzyskaliśmy z połączenia ESF($A3$)+VFH(Φ)+VFH(d) oraz ESF($A3$)+VFH(θ)+VFH(α)+VFH(Φ)+VFH(d). Uzyskane w ten sposób wyniki, odpowiednio 94.79% i 95.28%, są najlepsze wśród wszystkich przeprowadzonych testów walidacji krzyżowej. Połączenia cech zawierające histogramy GRSD okazały się znacznie słabsze niż najlepsze połączenia cech VFH i ESF. Warto jednak zauważyć, że GRSD wzmocnił wpływ VFH(d), podczas gdy osłabił ESF($A3$).

Ciekawym spostrzeżeniem jest fakt, że cecha ESF($A3$), czyli informacja na temat kątów pomiędzy liniami utworzonymi z trójek punktów, okazała się lepsza od informacji traktującej o zależnościach kątowych między normalnymi do powierzchni rozpiętej na chmurze punktów, którą reprezentuje *VFH(θ)+VFH(Φ)+VFH(α)*. Jednak w połączeniu z cechą VFH(d), określającą rozkład odległości pomiędzy poszczególnymi punktami, lepszą skuteczność klasyfikacji (przy takim samym rozmiarze wektora cech) otrzymujemy w przypadku cech osobnego deskryptora ESF niż tego samego deskryptora VFH.

Za punkt kompromisu pomiędzy skutecznością klasyfikacji, a ilością wymiarów danych (czyli rozmiarem wektora cech) można uznać kombinację cech ESF($A3$)+VFH(d) (skuteczność: 93.61%, rozmiar wektora: 24) lub ESF($A3$)+VFH(Φ)+VFH(d) (skuteczność: 94.79%, rozmiar wektora: 30). W przypadku pierwszej z wymienionych kombinacji, odchylenie standardowe z 30 testów walidacji krzyżowej wynosi 0.31%, natomiast dla drugiej kombinacji - 0.44%, a więc biorąc pod uwagę mniejszy rozrzut poprawności klasyfikacji, co może mieć znaczenie w przypadku stosowania metody w praktyce, pierwsza kombinacja wydaje się najlepszym kompromisem.

5. Podsumowanie

W artykule przedstawiono porównanie poszczególnych cech trzech deskryptorów chmur punktów w rozpoznawaniu statycznych układów dłoni. Omówiono ideę i algorytm wyznaczania każdego deskryptora, metodę rozpoznawania układów dłoni oraz sposób przeprowadzania eksperymentów. Zamieszczono wyniki testów walidacji krzyżowej w postaci tabelarycznej oraz wnioski i spostrzeżenia na ich temat. Główną konkluzją niniejszego artykułu jest stwierdzenie, że analizowane deskryptory chmur punktów nie muszą występować w całości oraz samodzielnie. Usuwanie poszczególnych cech danego de-

skryptora i łączenie ich z wybranymi cechami innych deskryptorów może się przyczynić do poprawy skuteczności klasyfikacji, co zostało sprawdzone na przykładzie chmur punktów reprezentujących statyczne układy dłoni. Ponadto, z przeprowadzonych eksperymentów można wywnioskować, że deskryptory VFH oraz ESF są bardziej odpowiednie dla rozpoznawania kształtów dłoni niż deskryptor GRSD.

Literatura

[1] Bay H., Ess A., Tuytelaars T., Van Gool L.: Speeded-Up Robust Features, Computer Vision and Image Understanding, tom 110, wydanie 3, czerwiec 2008, pp. 346-359.

[2] Bellman, R. E.: Adaptive Control Processes, Princeton, NJ. Press, 1961.

[3] Dominio F., Donadeo M., Zanuttigh P.: Combining multiple depth-based descriptors for hand gesture recognition, Pattern Recognition Letters, tom 50, grudzień 2014, pp. 101-111.

[4] Kapuscinski T., Oszust M., Wysocki M.: Recognition of signed dynamic expressions observed by ToF camera, Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA): pp. 291-296, Poznan 2013.

[5] Kapuscinski T., Oszust M., Wysocki M., Warchol D.: Recognition of hand gestures observed by depth cameras, International Journal of Advanced Robotic Systems, 2015.

[6] Keskin C., Kirac F., Kara Y. E., Akarun L.: Real time hand pose estimation using depth sensors, Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1228-1234, Barcelona 2011.

[7] Larose, D. T.: Discovering knowledge in data: an introduction to data mining, John Wiley & Sons, 2014, pp. 90-106.

[8] Marton Z.C., Pangeric D., Blodow N., Beetz M.: Combined 2D-3D categorization and classification for multimodal perception systems, International Journal of Robotics Research, tom 10, wydanie 11, wrzesień 2011, pp. 1378-1472.

[9] Marton Z.C., Pangeric D., Blodow N., Kleinehellefort J., Beetz M.: General 3D modelling of novel objects from a single view, Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pp. 3700-3705, Taipei 2010.

[10] Marton Z.C., Pangeric D., Rusu R. B., Holzbach A., Beetz M.: Hierarchical object categorization and appearance classification for mobile manipulation, Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on, pp. 365-370, Nashville 2010.

[11] Molina J., Escudero-Viñolo A., Signoriello A., Pardàs M., Ferràn C., Bescós J., Marqués F., Martínez J. M.: Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models, Machine Vision and Applications, tom 24, wydanie 1, styczeń 2013, pp. 187-204.

[12] Oprisescu S. R.: Automatic static hand gesture recognition using ToF cameras, European Signal Processing Conference, pp. 2748-2751, Bucharest 2012.
[13] Ren Z., Yuan J., Zhang Z.: Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera, MM '11 Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 1093-1096.
[14] Rusu R. B., Blodow N., Beetz M.: Fast point feature histograms (FPFH) for 3D registration, Robotics and Automation, 2009. ICRA '09. IEEE International Conference on, pp. 3212-3217, Kobe 2009.
[15] Rusu R. B., Bradski G., Thibaux R.: Fast 3D recognition and pose using the Viewpoint Feature Histogram, Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pp. 2155-2162, Taipei 2010.
[16] Rusu R. B., Cousins S.: 3D is here: Point Cloud Library (PCL), Robotics and Automation (ICRA), 2011 IEEE International Conference on, pp. 1-4, Shanghai 2011.
[17] Rusu R. B., Holzbach A., Beetz M..: Detecting and Segmenting Objects for Mobile Manipulation, Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 47-54, Kyoto 2009.
[18] Rusu R., Marton Z. C., Blodow N.: Learning informative point classes for the acquisition of object model maps, Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on, pp. 643-650, Hanoi 2008.
[19] Uebersax D., Gall J., Van den Bergh M.: Real-time sign language letter and word recognition from depth data, Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 383-390, Barcelona 2011.
[20] Van den Bergh M., Van Gool L.: Combining RGB and ToF cameras for real-time 3D hand gesture interaction, Applications of Computer Vision (WACV), 2011 IEEE Workshop on, pp. 66-72, Kona 2011.
[21] Wohlkinger W., Vincze M.: Ensemble of Shape Functions for 3D object classifications, Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on, pp. 2987-2992, Phuket 2011.
[22] http://kinectexplorer.blogspot.com (aktualizacja 13 grudzień 2011).

COMPARING AND COMBINING OF POINT CLOUD DESCRIPTORS' FEATURES IN STATIC HAND POSTURE RECOGNITION

Summary

The paper presents the method of recognizing static hand postures using three point cloud descriptors: Viewpoint Feature Histogram (VFH), Global Radius-Based Surface Descriptor (GRSD) and Ensemble of Shape Functions (ESF). These descriptors describe point clouds with the usage of the voxel grid, vectors normal to the cloud surface, distributions of the points' distances to their neighbors, and radii of spheres inscribed to parts of the surface. The cross-validation tests have been performed yielding the comparison of classification correctness for the single features, combined features of the same descriptor and of different descriptors. The tests have been per-

formed on a dataset containing 1000 depth maps: 10 different hand postures shown 10 times by 10 subjects. Before the feature extracting process, each point cloud must be preprocessed, including: segmentation (in order to separate the hand from the other cloud parts), rotation related to the hand center and the longest extended finger (in order to make the algorithm independent from the hand rotations around the axis perpendicular to the camera lens), and the points reduction (in order to make the calculations faster). The results are complemented by an additional information – the size of the feature vector used in the classification. It allows to find a combination of features that constitutes a point of compromise between the classification correctness and the number of data dimensions.

Keywords: point cloud descriptors, Viewpoint Feature Histogram, Global Radius-Based Surface Descriptor, Ensemble of Shape Functions

DOI: 10.7862/re.2016.5

Tekst złożono w redakcji: styczeń 2016
Przyjęto do druku: marzec 2016