

Piotr KOTAS, Grzegorz BOCEWICZ
Politechnika Koszalińska

AGREGACJA REPREZENTACJI CECH A GŁOSOWANIE LOKALNYCH DECYZJI W KOLEKTYWNYM ROZPOZNAWANIU OBIEKTÓW

Streszczenie. W artykule przedstawiono podejście do kolektywnego rozpoznawania obiektów oparte na rozproszonej architekturze wymiany i agregacji reprezentacji cech. Architektura obejmuje zespół współpracujących jednostek wizyjnych wyposażonych w spłotowe sieci neuronowe (CNN), które obserwują ten sam obiekt z różnych perspektyw. Każda jednostka przetwarza lokalny obraz i wyznacza reprezentację cech, następnie udostępnianą pozostałym jednostkom zespołu. Wymieniane reprezentacje są agregowane w celu utworzenia wspólnej reprezentacji obiektu stanowiącej podstawę decyzji klasyfikacyjnej. Proponowane podejście zweryfikowano eksperymentalnie z wykorzystaniem zmodyfikowanego zbioru COIL-100. Wyniki porównano z podejściem opartym na głosowaniu lokalnych decyzji klasyfikacyjnych.

AGGREGATION OF FEATURE REPRESENTATIONS VS. VOTING OF LOCAL DECISIONS IN COLLECTIVE OBJECT RECOGNITION

Summary. The paper presents an approach to collective object recognition based on a distributed architecture for exchanging and aggregating feature representations. The architecture comprises a team of collaborating vision units equipped with convolutional neural networks (CNNs) that observe the same object from different perspectives. Each unit processes a local image and extracts a feature representation, which is then shared with the remaining units of the team. The exchanged representations are aggregated to create a joint representation of the object, which serves as the basis for the classification decision. The proposed approach was experimentally verified using a modified COIL-100 dataset. The results were compared with an approach based on the voting of local classification decisions.

1. Wprowadzenie

Systemy zdalnego monitorowania coraz częściej wykorzystują przetwarzanie brzegowe i platformy mobilne, umożliwiające lokalną analizę danych [1, 19, 20]. Mimo że nadal dominują architektury scentralizowane, rozwój wbudowanych procesorów graficznych (GPU) i specjalizowanego sprzętu umożliwia efektywne wdrażanie spłotowych sieci neuronowych (CNN) w rozproszonych, heterogenicznych

urządzeniach [11, 12, 13, 21, 24]. Rozwiązania te ograniczają zależność systemu od łączności o dużej przepustowości oraz zwiększają jego odporność na zakłócenia komunikacyjne [3, 4]. Mogą również wspierać efektywne wykorzystanie istniejącej infrastruktury i planowanie jej rozwoju [6].

Rozproszone systemy percepcji wizyjnej są jednak ograniczone częściową obserwowalnością, ponieważ poszczególne urządzenia rejestrują jedynie fragmentaryczne informacje zależne od położenia względem obserwowanego obiektu. Istniejące rozwiązania opierają się głównie na agregacji lokalnych decyzji klasyfikacyjnych, np. głosowaniu nad wynikami klasyfikacji [8, 10]. Podejścia te redukują obserwacje jednostek do lokalnych decyzji, przez co nie wykorzystują w pełni komplementarnych informacji przestrzennych dostępnych w różnych obserwacjach.

W artykule rozważono współpracę jednostek roju wizyjnego realizowaną poprzez wymianę i agregację pośrednich reprezentacji cech. Zamiast wymieniać końcowe decyzje klasyfikacyjne, jednostki udostępniają reprezentacje cech wyznaczone przez sieci CNN. Reprezentacje te są następnie agregowane przed klasyfikacją w celu utworzenia wspólnej reprezentacji obserwowanego obiektu. W odróżnieniu od wcześniejszych badań, w tym [9], zakładających identyczne lub statystycznie niezależne dane wejściowe [15, 16], rozważono obserwacje tego samego obiektu pozyskiwane z różnych perspektyw. W celu wyizolowania wpływu mechanizmu agregacji pominięto ograniczenia komunikacyjne, takie jak opóźnienia, utrata danych i synchronizacja transmisji.

Artykuł przedstawia rozproszony model rozpoznawania wizualnego, w którym decyzja kolektywna jest wyznaczana na podstawie agregacji reprezentacji cech. Zaproponowane podejście zweryfikowano na zmodyfikowanym zbiorze COIL-100 [14], rozszerzonym o zróżnicowane tekstury tła oraz zmiany skali obiektu odpowiadające różnym odległościom obserwacji.

2. Przegląd literatury

Rozwój systemów zdalnego monitorowania wizyjnego wynika z potrzeby równoważenia możliwości obliczeniowych i zakresu transmisji danych. W tradycyjnych architekturach, stosowanych m.in. w mobilnych systemach czujnikowych oraz platformach UAV, urządzenia brzegowe przesyłają surowe dane (najczęściej są to pozyskiwane obrazy obserwowanego obiektu) do systemów chmurowych lub stacji naziemnych, w których są one analizowane [19, 20]. Rozwiązanie to umożliwia wykorzystanie złożonych modeli, lecz wymaga niezawodnej łączności o dużej przepustowości. Ograniczenie to może prowadzić do wzrostu opóźnień i obniżenia niezawodności systemu, szczególnie w zastosowaniach krytycznych, takich jak reagowanie kryzysowe lub rolnictwo precyzyjne [3, 4, 17]. Architektury hierarchiczne częściowo ograniczają te problemy, jednak nadal zakładają centralną koordynację i ograniczoną autonomię urządzeń.

Rozwój wbudowanych procesorów graficznych i specjalizowanego sprzętu umożliwił realizację wnioskowania z wykorzystaniem sieci neuronowych bezpośrednio na urządzeniach brzegowych [11, 13]. Rozproszenie obliczeń zwiększa skalowalność i odporność systemu, lecz nie eliminuje ograniczeń wynikających z częściowej obserwowalności. Każda jednostka dysponuje bowiem jedynie fragmentaryczną

informacją o obserwowanym obiekcie, zależną od jej położenia i warunków obserwacji. W systemach wieloagentowych integracja informacji jest zwykle realizowana z wykorzystaniem mechanizmów konsensusu lub wymiany informacji lokalnych [15, 16]. Skuteczne łączenie heterogenicznych obserwacji bez centralnej koordynacji pozostaje jednak problemem otwartym.

W rozproszonych systemach rozpoznawania najczęściej stosuje się agregację lokalnych decyzji, realizowaną np. przez głosowanie twarde lub miękkie [8, 10]. Podejścia te są proste obliczeniowo, lecz redukują każdą obserwację do końcowej decyzji klasyfikacyjnej. W konsekwencji nie wykorzystują w pełni informacji przestrzennych i kontekstowych zawartych w pośrednich reprezentacjach cech, zwłaszcza gdy lokalne obserwacje są częściowo zasłonięte lub zdegradowane.

Alternatywę stanowi wymiana reprezentacji cech wyznaczanych przez sieci CNN przed wykonaniem klasyfikacji. Agregacja reprezentacji cech umożliwia łączenie informacji zachowanych w lokalnych reprezentacjach i może zapewniać skuteczniejszą integrację obserwacji niż głosowanie lokalnych decyzji klasyfikacyjnych. Dotychczasowe badania wskazują na jej przewagę w środowiskach multimodalnych oraz przy zdegradowanych danych wejściowych [2, 23]. Rozważane w nich dane wejściowe są jednak zazwyczaj identyczne lub statystycznie niezależne, przez co nie uwzględniają wprost komplementarności obserwacji wynikającej z różnych perspektyw.

Rozpoznawanie obiektów obserwowanych z różnych perspektyw było dotychczas analizowane głównie w systemach scentralizowanych, w których wszystkie obrazy są dostępne dla wspólnego modelu [18]. W architekturach rozproszonych wspólna reprezentacja nie jest dostępna bez wymiany informacji między jednostkami. Poszczególne obserwacje mogą przy tym dostarczać komplementarnych, a nie jedynie redundantnych danych o tym samym obiekcie. Powstaje zatem potrzeba analizy mechanizmów umożliwiających agregację takich obserwacji w systemach działających przy częściowej obserwowalności.

W niniejszej pracy rozważono rozproszoną architekturę kolektywnego rozpoznawania obiektów opartą na wymianie i agregacji reprezentacji cech. Analiza koncentruje się na ocenie, czy agregacja reprezentacji cech umożliwia utworzenie wspólnej reprezentacji zapewniającej skuteczną klasyfikację obiektu obserwowanego z różnych perspektyw.

W następnej sekcji przedstawiono formalne sformułowanie rozważanego problemu, obejmujące porównanie agregacji reprezentacji cech z głosowaniem lokalnych decyzji klasyfikacyjnych.

3. Sformułowanie problemu

Rozważany jest rój U złożony z N jednostek obserwujących ten sam obiekt O : $U = \{u_1, u_2, \dots, u_i, \dots, u_N\}$. Jednostki realizują rozproszone zadanie rozpoznawania obiektów. Ze względu na różne położenia względem obserwowanego obiektu każda jednostka u_i uzyskuje lokalny obraz x_i , odpowiadający określonej konfiguracji obserwacji:

$$x_i = h(O, \theta_i, d_i) \quad (1)$$

gdzie θ_i oznacza kąt obserwacji, natomiast d_i – odległość jednostki u_i od obiektu O .

Lokalne obrazy mogą różnić się zakresem widocznych fragmentów obiektu, jego skalą, charakterystyką tła oraz stopniem przesłonięcia. Każda jednostka przetwarza lokalny obraz z wykorzystaniem własnej funkcji ekstrakcji cech f_{ϕ_i} . Dla obrazu x_i wyznaczana jest reprezentacja:

$$z_i = f_{\phi_i}(x_i) \in \mathbb{R}^m \quad (2)$$

Funkcje ekstrakcji cech stosowane w poszczególnych jednostkach mogą różnić się architekturą lub parametrami, jednak są trenowane dla tego samego zadania klasyfikacyjnego. Przyjmuje się ich tożsamość klasyfikacyjną, rozumianą jako możliwość poprawnego przypisania obserwacji tego samego obiektu do tej samej klasy na podstawie wyznaczanych reprezentacji cech. Przyjmuje się ponadto, że wszystkie reprezentacje z_i należą do wspólnej m -wymiarowej przestrzeni cech i mogą podlegać agregacji.

W rozważanej architekturze lokalne reprezentacje cech są wymieniane między jednostkami i agregowane przed wykonaniem klasyfikacji. Niech $Z = \{z_1, z_2, \dots, z_i, \dots, z_N\}$ oznacza zbiór reprezentacji cech wyznaczonych przez jednostki roju. Reprezentację zagregowaną określa operator Φ :

$$z^* = \Phi(z_1, z_2, \dots, z_i, \dots, z_N). \quad (3)$$

Reprezentacja z^* integruje informacje uzyskane przez poszczególne jednostki i stanowi podstawę kolektywnego rozpoznawania obiektu. Decyzja kolektywna jest wyznaczana przez funkcję klasyfikacyjną g_c :

$$\hat{y}^* = g_c(z^*) \quad (4)$$

gdzie $\hat{y}^*$ oznacza przewidywaną klasę obserwowanego obiektu.

Dla porównania, w podejściu opartym na głosowaniu lokalnych decyzji klasyfikacyjnych każda jednostka wyznacza własną decyzję klasyfikacyjną: $\hat{y}_i = g_i(z_i)$ gdzie g_i oznacza lokalną funkcję klasyfikacyjną jednostki u_i . Następnie decyzja kolektywna jest wyznaczana z wykorzystaniem operatora głosowania Ψ :

$$\hat{y}^V = \Psi(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N). \quad (5)$$

Różnica między rozważanymi podejściami dotyczy etapu integracji informacji. W architekturze wymiany cech agregacja jest wykonywana przed klasyfikacją, na poziomie pośrednich reprezentacji. W podejściu opartym na głosowaniu integracji podlegają wyłącznie lokalne decyzje klasyfikacyjne. Redukcja lokalnej obserwacji do pojedynczej decyzji może prowadzić do utraty informacji zawartej w reprezentacji cech.

Rozważany problem polega na ocenie, czy dla roju U , którego jednostki obserwują ten sam obiekt O , agregacja reprezentacji cech zapewnia większą jakość rozpoznawania niż agregacja lokalnych decyzji klasyfikacyjnych. Rozproszony model rozpoznawania wyznacza decyzję:

$$\hat{y}^* = g_c(\Phi(z_1, z_2, \dots, z_i, \dots, z_N)) \quad (6)$$

Jakość rozpoznawania ocenia się na podstawie dokładności klasyfikacji. Przyjęty próg τ określa minimalną wartość dokładności wymaganą dla rozważanego scenariusza:

$$\text{Acc}(F, \Phi, g_c) \geq \tau \quad (7)$$

gdzie $F = \{f_{\phi_1}, f_{\phi_2}, \dots, f_{\phi_N}\}$ oznacza zbiór funkcji ekstrakcji cech wykorzystywanych przez jednostki roju, $\text{Acc}(F, \Phi, g_c)$ oznacza dokładność klasyfikacji, a τ jest przyjętym minimalnym progiem dokładności.

Celem prowadzonych badań jest:

1. ocena dokładności klasyfikacji architektury wykorzystującej wymianę i agregację reprezentacji cech;
2. porównanie jej dokładności klasyfikacji z podejściem opartym na głosowaniu lokalnych decyzji klasyfikacyjnych;
3. analiza wpływu zróżnicowania obserwacji, obejmującego kąty obserwacji θ_i , odległości d_i oraz częściową obserwowalność, na dokładność klasyfikacji.

W analizie przyjęto idealne warunki komunikacji. Reprezentacje cech są przesyłane bez opóźnień, strat i ograniczeń przepustowości. Założenie to umożliwia ocenę wpływu mechanizmu wymiany i agregacji cech, natomiast analiza ograniczeń komunikacyjnych pozostaje przedmiotem dalszych badań.

4. Architektura systemu

Zgodnie z ogólną koncepcją przedstawioną w [9] rozważaną architekturę kolektywnego rozpoznawania obiektów (rys. 1) można opisać krotką:

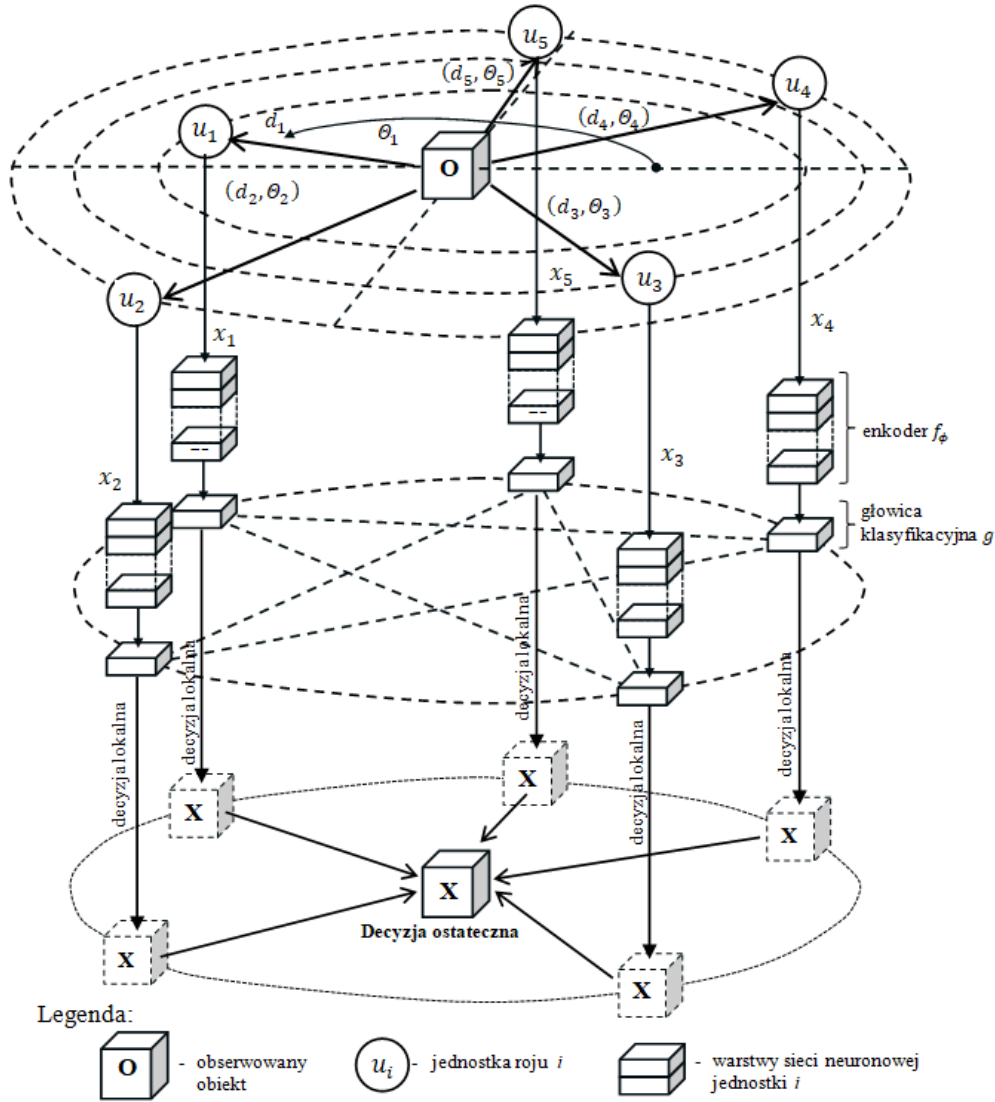
$$M = (U, X, Z, F, \Phi, g_c), \quad (8)$$

gdzie: $U = \{u_1, u_2, \dots, u_N\}$ oznacza rój jednostek, $X = \{x_1, x_2, \dots, x_N\}$ jest zbiorem lokalnych obrazów, $Z = \{z_1, z_2, \dots, z_N\}$ oznacza zbiór reprezentacji cech wyznaczonych przez jednostki roju, natomiast $F = \{f_{\phi_1}, f_{\phi_2}, \dots, f_{\phi_N}\}$ jest zbiorem funkcji ekstrakcji cech. Operator Φ określa mechanizm agregacji reprezentacji, a g_c oznacza funkcję klasyfikacyjną stosowaną do reprezentacji zagregowanej.

Każdy obraz lokalny x_i przedstawia ten sam obiekt O , obserwowany przez jednostkę u_i z określonego kąta θ_i i odległości d_i . Poszczególne obserwacje mogą zatem różnić się widocznymi fragmentami obiektu, jego skalą, tłem oraz stopniem przesłonięcia.

Dla każdej jednostki u_i lokalny obraz x_i jest przekształcany w reprezentację cech: $z_i = f_{\phi_i}(x_i)$, gdzie: $f_{\phi_i}: \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^m$. Symbole H , W i C oznaczają odpowiednio wysokość, szerokość i liczbę kanałów obrazu wejściowego, natomiast m określa wymiar przestrzeni cech. W rozważanej implementacji zastosowano sieć ResNet50 [5]. Obrazy wejściowe mają postać RGB o rozmiarze 224×224 , a zatem $C=3$. W konsekwencji funkcje ekstrakcji cech mają postać: $f_{\phi_i}: \mathbb{R}^{224 \times 224 \times 3} \rightarrow \mathbb{R}^m$.

Funkcje ekstrakcji cech mogą różnić się parametrami, a w ogólnym przypadku również architekturą. Przyjmuje się jednak ich tożsamość klasyfikacyjną oraz możliwość wyznaczania reprezentacji należących do wspólnej przestrzeni cech $\mathbb{R}^m$. Umożliwia to agregację reprezentacji wyznaczonych przez różne jednostki roju.



Rys. 1 Schemat proponowanej architektury

Zbiór lokalnych reprezentacji cech jest następnie agregowany z wykorzystaniem operatora Φ . W rozważanym modelu zastosowano uśrednianie:

$$z^* = \Phi(z_1, z_2, \dots, z_N) = \frac{1}{N} \sum_{u_i \in U} z_i. \quad (9)$$

Ponieważ: $z_i \in \mathbb{R}^m$, zagregowana reprezentacja również spełnia warunek: $z^* \in \mathbb{R}^m$. Reprezentacja z^* integruje informacje uzyskane przez wszystkie jednostki obserwujące ten sam obiekt. Uśrednianie reprezentacji może ograniczać wpływ lokalnych zakłóceń oraz łączyć cechy pochodzące z różnych perspektyw obserwacji.

Ostateczna decyzja klasyfikacyjna jest wyznaczana przez funkcję klasyfikacyjną g_c , która przypisuje zagregowanej reprezentacji rozkład prawdopodobieństwa dla K klas: $g_c(z^*) = (g_1(z^*), g_2(z^*), \dots, g_K(z^*)) \in \Delta_K$, gdzie: $\Delta_K = \{q \in \mathbb{R}^K : q_k \geq 0, \sum_{k=1}^K q_k = 1\}$. Przewidywana etykieta klasy jest określana następująco:

$$\hat{y}^* = \arg \max_{k \in \{1, \dots, K\}} g_k(z^*). \quad (10)$$

Wartość $g_k(z^*)$ oznacza prawdopodobieństwo przypisane przez klasyfikator do klasy k . Decyzja kolektywna jest zatem wyznaczana na podstawie zagregowanej reprezentacji cech, a nie przez połączenie niezależnych lokalnych decyzji klasyfikacyjnych. Dla tożsamej głowicy klasyfikacji g_c decyzja kolektywna zachodzi równolegle, tzn. każda jednostka dochodzi do tego samego wyniku i działa na tej podstawie.

Działanie architektury można podsumować następująco:

1. Lokalne wykrycie: $X = \{x_i: u_i \in U\}, x_i = h(O, \theta_i, d_i) \in \mathbb{R}^{H \times W \times C}$
2. Kodowanie cech: $z_i = f_{\phi_i}(x_i) \in \mathbb{R}^m, Z = \{z_i: u_i \in U\}$
3. Agregacja cech: $z^* = \Phi(z_1, z_2, \dots, z_N) = \frac{1}{N} \sum_{u_j \in U} z_j$
4. Klasyfikacja: $g(z^*) = (g_1(z^*), g_2(z^*), \dots, g_K(z^*)) \in \Delta^K$
5. Decyzja: $\hat{y} = \arg \max_{k \in \{1, \dots, K\}} g_k(z^*)$

5. Eksperymenty obliczeniowe

Celem eksperymentów była ocena dokładności klasyfikacji zaproponowanej architektury oraz jej porównanie z podejściem opartym na głosowaniu decyzji klasyfikacyjnych. Analizowano wpływ zróżnicowania perspektyw obserwacji, odległości od obiektu oraz częściowego przesłonięcia obrazu.

Do badań wykorzystano zbiór COIL-100, zawierający obrazy 100 obiektów rejestrowanych pod różnymi kątami. Analizę przeprowadzono dla rojów o liczebności: $N \in \{1, 2, \dots, 10\}$. Każdej jednostce u_i przypisywano obraz obiektu zarejestrowany dla kąta obserwacji: $\theta_i \in \{0^\circ, 5^\circ, \dots, 355^\circ\}$. Kąty obserwacji dobierano tak, aby nie powtarzały się między jednostkami roju. Przyjęto, że parametr d_i opisujący odległość obserwacji przyjmuje wartości: $d_i \in \{0,8; 1,0; 1,33; 2,0\}$, gdzie $d_i = 1,0$ odpowiada odległości nominalnej, tj. skali obiektu przyjętej podczas treningu. Wartość $d_i = 2,0$ odpowiada dwukrotnie większej odległości obserwacji, a tym samym mniejszej skali obiektu w obrazie.

W celu uwzględnienia zróżnicowania warunków obserwacji obrazy obiektów umieszczono na pięciu teksturach tła wykorzystywanych podczas treningu. Szóstą teksturę wykorzystano wyłącznie do utworzenia zbioru testowego. Przykładowe obrazy przedstawiono na rysunku 2. Przyjęto minimalny próg dokładności rozpoznawania: $\tau = 0,90$.



Rys. 2 Przykład zmodyfikowanych obiektów COIL-100 na podłożach z drewna, mchu i piasku.

Każdy obraz oceniano w dwóch wariantach: bez degradacji oraz z częściowym przesłonięciem. Wariant przesłonięty uzyskano przez zakrycie 25% powierzchni obrazu czarnym kwadratem. Nie uwzględniono innych rodzajów degradacji, takich jak rozmycie gaussowskie, rozmycie ruchu ani szum impulsowy – będą one przedmiotem dalszych badań.

Funkcje ekstrakcji cech f_ϕ oraz funkcje klasyfikacyjne g wykorzystane przez jednostki roju wytrenowano z zastosowaniem optymalizatora Adam [7], przy współczynniku uczenia 0,001 i rozmiarze partii 16. Zbiór treningowy utworzono przez rozszerzenie zbioru COIL-100 o pięć tekstur tła; obejmował on 36 000 obrazów. Zbiór testowy skonstruowano z wykorzystaniem odrębnej tekstury tła oraz pięciu wariantów odległości obserwacji. Dla każdego poziomu odległości przygotowano 7200 obrazów testowych, co łącznie dało 28 800 obrazów. Modele optymalizowano z wykorzystaniem funkcji straty entropii krzyżowej (Cross-Entropy). Punkty kontrolne zapisywano po każdej epoce treningu.

Scenariusze eksperymentalne skonstruowano tak, aby wszystkie jednostki roju obserwowały ten sam obiekt przy ustalonej wartości parametru odległości: $d_i = d, \forall u_i \in U$. Lokalne obrazy różniły się wyłącznie kątem obserwacji. Każdy scenariusz oceniano dla $N = 1, \dots, 10$, w wariancie bez przesłonięcia oraz z częściowym przesłonięciem obrazu. Jakość rozpoznawania określono za pomocą dokładności klasyfikacji:

$$\text{Acc} = \frac{n_{\text{corr}}}{n_{\text{tot}}}, \quad (11)$$

gdzie: n_{corr} – oznacza liczbę poprawnie sklasyfikowanych instancji, a n_{tot} – całkowitą liczbę analizowanych instancji.

W analizie przyjęto idealne warunki komunikacji. Reprezentacje cech były przesyłane bez opóźnień, utraty danych i ograniczeń przepustowości. Ponadto w każdym scenariuszu wszystkie jednostki podlegały temu samemu rodzajowi i poziomowi degradacji obrazu. Łącznie oceniono 50 000 instancji.

Zagregowane wyniki przedstawiono na rysunku 3. Wskazują one, że jakość rozpoznawania zależy przede wszystkim od odległości obserwacji i obecności częściowego przesłonięcia. Dla $d_i = 0,8$ oraz $d_i = 1,0$ oba podejścia osiągają wysoką dokładność już dla niewielkiej liczby jednostek. Dokładność klasyfikacji zbliża się do 100%, również w wariancie z częściowym przesłonięciem. Próg $\tau = 0,90$ jest w tych warunkach osiągany już dla pojedynczej jednostki i utrzymuje się dla wszystkich analizowanych liczebności roju.

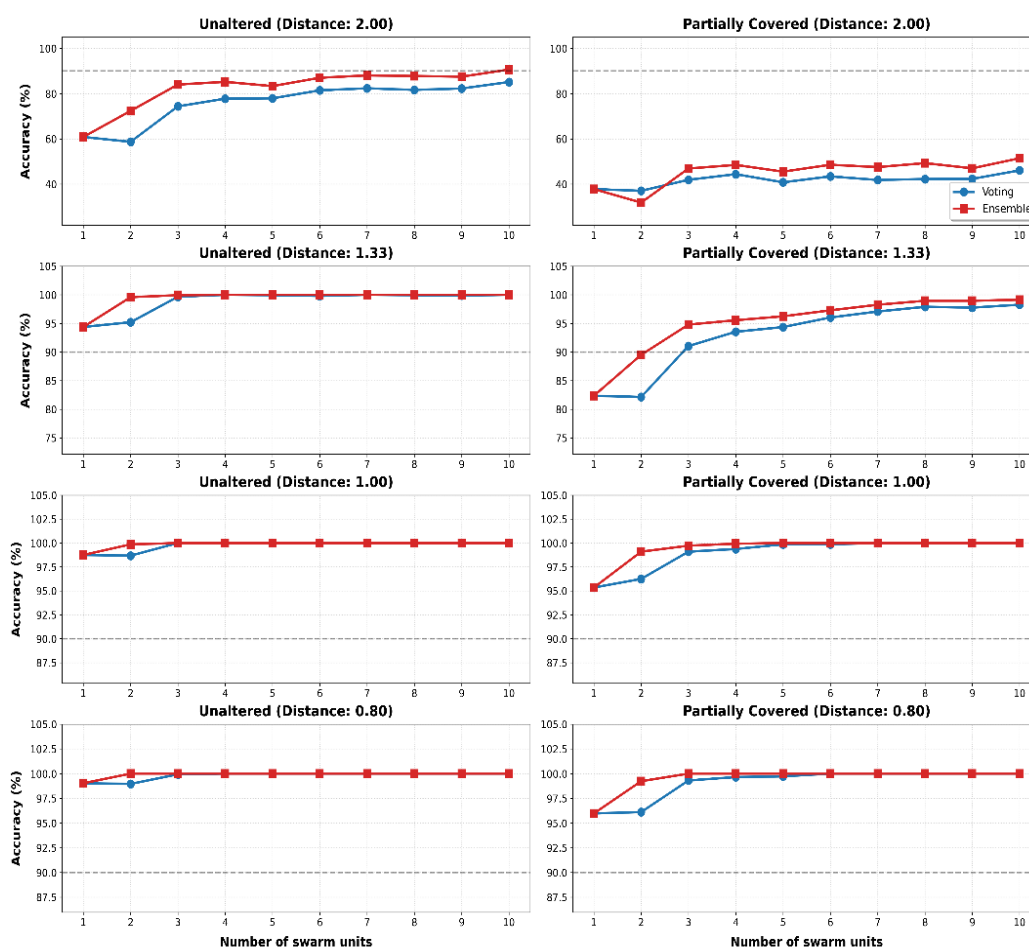
Dla $d_i = 1,33$ w wariancie bez przesłonięcia oba podejścia nadal osiągają niemal maksymalną dokładność klasyfikacji. W przypadku częściowego przesłonięcia przewaga agregacji reprezentacji cech staje się bardziej widoczna. Dla większości liczebności roju agregacja reprezentacji pośrednich zapewnia wyższą dokładność niż głosowanie, przy zachowaniu progu $\tau = 0,90$.

Największe różnice występują dla $d_i = 2,0$ (dwukrotne zwiększenie odległości względem odległości nominalnej). W przypadku obrazów bez przesłonięcia dokładność klasyfikacji obu podejść wzrasta wraz z liczebnością roju, jednak agregacja reprezentacji cech osiąga na ogół wyższe wartości niż głosowanie. Przewaga ta jest wyraźniejsza w wariancie częściowego przesłonięcia. Dla $d_i = 2,0$ i obrazów

przesłoniętych końcowa dokładność wynosi 51,4% dla agregacji reprezentacji cech oraz 46,0% dla głosowania.

Uzyskane wyniki wskazują, że agregacja reprezentacji cech jest szczególnie korzystna w warunkach, w których lokalne obserwacje pozostają informacyjne, lecz są częściowo niekompletne wskutek zwiększonej odległości obserwacji lub przesłonięcia. Dla $d_i = 0,8$, $d_i = 1,0$ i $d_i = 1,33$, próg $\tau = 0,90$ jest osiągalny, natomiast dla $d_i = 2,0$, szczególnie w wariancie z przesłonięciem, nie zostaje utrzymany.

Uzyskane wyniki wskazują, że agregacja reprezentacji cech jest szczególnie korzystna, gdy lokalne obrazy zachowują cechy dyskryminacyjne umożliwiające rozróżnienie obiektów, lecz zawierają niepełną informację wskutek zwiększonej odległości obserwacji lub przesłonięcia. Dla $d_i = 0,8$, $d_i = 1,0$ i $d_i = 1,33$ próg $\tau = 0,90$ jest osiągalny, natomiast dla $d_i = 2,0$, szczególnie w wariancie z przesłonięciem, nie zostaje utrzymany.



Rys. 3 Dokładność klasyfikacji linii bazowej głosowania i proponowanego podejścia agregacji reprezentacji cech dla obrazów niezmienionych (kolumna 1) i obrazów częściowo pokrytych (kolumna 2) dla odległości $d_i = 2,0, 1,33, 1,0, 0,8$

6. Wnioski

W artykule przedstawiono rozproszoną architekturę kolektywnego rozpoznawania obiektów opartą na wymianie i agregacji reprezentacji cech. Jej dokładność klasyfikacji

porównano z podejściem wykorzystującym głosowanie lokalnych decyzji klasyfikacyjnych. Wyniki wskazują, że agregacja reprezentacji cech może zwiększać dokładność rozpoznawania, gdy lokalne obrazy dostarczają komplementarnych informacji o tym samym obiekcie.

Eksperymenty przeprowadzone na zmodyfikowanym zbiorze COIL-100 potwierdziły wpływ odległości obserwacji, częściowego przesłonięcia oraz charakterystyki tła na dokładność klasyfikacji. W warunkach korzystnej obserwacji oba porównywane podejścia osiągają wysoką dokładność. Przewaga agregacji cech jest najbardziej widoczna, gdy lokalne obrazy zachowują cechy dyskryminacyjne umożliwiające rozróżnienie obiektów, lecz zawierają niepełną informację wskutek zwiększonej odległości obserwacji lub przesłonięcia.

W najbardziej wymagających scenariuszach oba podejścia wykazują ograniczoną dokładność klasyfikacji. Gdy obiekt zajmuje niewielką część obrazu, lokalne reprezentacje nie zawierają wystarczającej informacji różnicującej, a agregacja cech nie zapewnia osiągnięcia przyjętego progu dokładności. Wyniki wskazują również, że złożone tło może obniżać jakość reprezentacji cech i ograniczać korzyści wynikające z ich agregacji.

Przedstawione wyniki mają charakter empiryczny. Nie określono formalnych warunków, w których agregacja reprezentacji cech zapewnia przewagę nad agregacją lokalnych decyzji. Dalsze badania powinny koncentrować się więc na określeniu warunków korzystnej agregacji oraz zostać rozszerzone o eksperymenty uwzględniające m.in. sceny zawierające wiele obiektów, dodatkowe rodzaje degradacji obrazu, w tym rozmycie i szum impulsowy, a także bardziej zaawansowane mechanizmy agregacji. Konieczne jest również uwzględnienie ograniczeń komunikacyjnych, takich jak opóźnienia, utrata danych i ograniczona przepustowość.

LITERATURA

1. Attalah M., Zaidi S., Mellal N., Calafate C.: Task-Offloading Optimization Using a Genetic Algorithm in Hybrid Fog Computing for the Internet of Drones. *Sensors* 2025, 25(5), <https://doi.org/10.3390/s25051383>
2. Feichtenhofer C., Pinz A., Zisserman A.: Convolutional Two-Stream Network Fusion for Video Action Recognition, 2016, pp.1933-1941, 10.1109/CVPR.2016.213
3. Gupta L., Jain R., Vaszkun G.: Survey of Important Issues in UAV Communication Networks. *IEEE Communications Surveys&Tutorials*. 18, 2015, 1-1. 10.1109/COMST.2015.2495297
4. Hayat S., Yanmaz E., Muzaffar R.: Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. *IEEE Communications Surveys&Tutorials*, 2016, 18. 1-1. 10.1109/COMST.2016.2560343
5. He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition, 2015, <https://doi.org/10.48550/arXiv.1512.03385>
6. Kashyap R., Bhuvan M.S., Chamarti S., Bhat P., Jothish M., Annappa K.: Algorithmic Approach for Strategic Cell Tower Placement, <https://doi.org/10.1109/ISMS.2014.112>, 2014.
7. Kingma D. P., Ba J.: Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego, 2015, <https://doi.org/10.48550/arXiv.1412.6980>

8. Kittler J., Hatef M., Duin R. P. W., Matas J.: On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(3), p.226-239. <https://doi.org/10.1109/34.667881>
9. Kotas P., Bocewicz G., Banaszak Z.: Evaluating CNN FeatureSharing in Multi-Device Visual Monitoring. *AUTOMATION 2025. Lecture Notes in Networks and Systems*, in print
10. Kuncheva L.: *Combining Pattern Classifiers*. Wiley, 2004, DOI:10.1002/0471660264
11. Lane N., Bhattacharya S., Georgiev P., Forlivesi C., Jiao L., Qendro L., Kawsar F.: DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices, 2016, 1-12. 10.1109/IPSAN.2016.7460664
12. Mandloi D., Arya R., Verma A.: Internet of Drones, 2024, https://doi.org/10.1007/978-981-97-6790-8_13
13. Mittal S.: A survey of techniques for improving energy efficiency in embedded computing systems. *International Journal of ComputerAided Engineering and Technology*, 2013, 6. 10.1504/IJCAET.2014.065419
14. Murase H., Nayar S. K., Nene S. A.: Columbia object image library (COIL-100) 1996, (Technical Report CUCS-006-96)
15. Neverova N., Wolf C., Taylor G., Nebout F.: ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 38. 10.1109/TPAMI.2015.2461544
16. Olfati-Saber R., Fax J., Murray R.: Consensus and Cooperation in Networked Multi-Agent Systems, 2007, *Proceedings of the IEEE*. 95. 215-233. 10.1109/JPROC.2006.887293.
17. Postawka A., Rudy J.: Lifelogging system based on Averaged Hidden Markov Models: Dangerous activities recognition for care giver support. *Computer Science*, 19 (3), pp. 257 - 279, 2018, DOI: 10.7494/csci.2018.19.3.2855
18. Qi C., R., Su H., Mo K., Guibas L.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, 2016, 10.48550/arXiv.1612.00593.
19. Satyanarayanan, M.: The Emergence of Edge Computing. *Computer*, 2017, 50. 30-39. 10.1109/MC.2017.9
20. Shi W., Cao J., Zhang Q., Li Y., Xu L.: Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 2016, 3. 1-1. 10.1109/JIOT.2016.2579198
21. Silva-Neto H. A., Sousa D. S., Duarte L. C., Petrucci J. F. S., Coltro W. K. T.: Recent developments on aerial lab-on-a-drone platforms for remote environmental monitoring: A review, 2025, <https://doi.org/10.1016/j.aca.2024.343544>
22. Su H., Maji S., Kalogerakis E., Learned-Miller, E.: Multi-view Convolutional Neural Networks for 3D Shape Recognition, 2015, 10.1109/ICCV.2015.114
23. Wang J., Hu X.: Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications (Ren, W. and Beard, R.W.; 2008) *IEEE Control Systems Magazine – IEEE CONTROL SYST MAG*, 2010, 30. 85-86. 10.1109/MCS.2010.936430
24. Yadav V. K. G., Kumar B.: Lightweight Surveillance Image Classification Through Hardware-Software Co-Design. *IEEE Embedded Systems Letters*, 2025, vol. 17, <https://doi.org/10.1109/LES.2025.3534237>