Politechnika Wrocławska

RADA DYSCYPLINY NAUKOWEJ
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA

ROZPRAWA DOKTORSKA

Optymalizacja procesu klasyfikacji dynamicznych scen bokserskich przez segmentację obrazu

mgr Piotr Stefański

Promotor
dr hab. Jan Kozak, prof. UE

Promotor pomocniczy
dr Tomasz Jach

WROCŁAW 2024

Spis treści

Wstęp

Obiektywy aparatów oraz kamer generują coraz większe ilości danych, których nie sposób analizować manualnie. Zatem niezbędne są rozwiązania, które w sposób automatyczny będą dostarczały wartościowych informacji na temat analizowanego obrazu. Taka problematyka jest w zakresie dziedziny wizji komputerowej (ang. computer vision, CV), w której naukowcy opracowali szereg algorytmów do klasyfikacji rejestrowanej sceny.

Automatyczna analiza wideo w sportach zyskała na znaczeniu w ostatnich latach, stając się kluczowym narzędziem dla trenerów, analityków oraz samych sportowców. Technologia ta umożliwia szczegółową analizę zachowań zawodników, strategii gry oraz wydajności fizycznej. W kontekście boksu, gdzie dynamika ruchów i szybkość reakcji odgrywają kluczową rolę, analiza wideo pozwala na dokładne zrozumienie technik zawodników, ich słabych punktów oraz potencjalnych obszarów do poprawy. Dzięki zaawansowanym algorytmom wizji komputerowej możliwe jest automatyczne wykrywanie i klasyfikowanie specyficznych zdarzeń na ringu, co stanowi istotne wsparcie dla decyzji trenerskich. Pomimo znaczących postępów w tej dziedzinie, nadal istnieją wyzwania związane z analizą scen, gdzie kluczowe obiekty zajmują niewielką część obrazu, co wymaga dalszych badań i optymalizacji procesów segmentacji oraz klasyfikacji dynamicznych scen bokserskich.

Klasyfikacja obrazów, na których obiekty determinujące klasę decyzyjną zajmują jedynie niewielki procent całkowitej powierzchni obrazu, stanowi poważne wyzwanie w dziedzinie wizji komputerowej. Problem ten jest szczególnie widoczny w kontekście sportów, takich jak boks, gdzie kluczowe dla analizy obiekty, takie jak rękawice zawodników, mogą zajmować zaledwie kilka procent powierzchni całego kadru. W takich przypadkach tradycyjne algorytmy klasyfikacji obrazu stają się nieefektywne. Niski stosunek powierzchni obiektów determinujących klasę decyzyjną do tła, które stanowi szum informacyjny skutecznie utrudnia poprawną klasyfikację, co następnie prowadzi do błędnych wniosków i nieefektywnych decyzji opartych na analizie wideo.

Rozwiązanie problemu klasyfikacji obrazów zawierających istotne obiekty na niewielkiej powierzchni jest aktualne oraz niezwykle potrzebne. Segmentacja obrazu, jako technika precyzyjnego wydzielania informacyjnych obiektów od tła, stanowi kluczowy element poprawy jakości klasyfikacji w takich scenach. Dzięki segmentacji możliwe jest wyodrębnienie i skoncentrowanie uwagi algorytmu na najważniejszych elementach obrazu, co prowadzi do zwiększenia dokładności analizy. W kontekście sportów, takich jak boks, segmentacja umożliwia szczegółowe monitorowanie i analizę ruchów zawodników, co jest nieocenione dla trenerów i analityków sportowych. Zastosowanie zaawansowanych technik segmentacji przed klasyfikacją może znacząco poprawić jakość wyników, umożliwiając bardziej precyzyjne wnioskowanie.

Dlatego w celu poprawy jakości klasyfikacji klatek wideo walki bokserskiej, na których znaczące obiekty zajmują bardzo małą powierzchnię rejestrowanej sceny, zaproponowane zostanie podejście do segmentacji klatek przed ich klasyfikacją. Podejście zostanie sprawdzone eksperymentalnie oraz porównane z podejściem bazowym oraz z innymi podejściami z literatury.

Teza

Zastosowanie dedykowanych algorytmów przetwarzania obrazów i uczenia maszynowego w podejściu do segmentacji obrazu umożliwia skrócenie czasu przetwarzania danych, przy jednoczesnym utrzymaniu wysokiego poziomu wydajności oraz stabilności klasyfikacji klatek wideo, na których istotne obiekty zajmują bardzo małą powierzchnię rejestrowanej sceny w problemie liczenia ciosów w walkach bokserskich.

Cel pracy

Głównym celem pracy jest opracowanie rozwiązania do segmentacji obrazu, które znacząco skróci czas przetwarzania danych w problemie liczenia ciosów w walkach bokserskich. Zaproponowane podejście jednocześnie musi utrzymać wysoki poziom wydajności i stabilności procesu klasyfikacji podczas pracy na obrazach, na których znaczące obiekty o charakterze informacyjnym zajmują jedynie poniżej 1,5 % powierzchni.

Cele poboczne

Głównym celem pobocznym pracy jest zbudowanie własnej bazy danych walk bokserskich jako elementarnej części procesu KDD (ang. knowledge discovery in databases, KDD). W tym celu należy dobrać odpowiedni sprzęt, wybrać odpowiednie wydarzenie sportowe, uzyskać zgodę organizatora, a następnie nagrać zmagania zawodników. Wynikiem tego etapu prac będzie zebranie blisko 500 GB materiału filmowego. Pozyskany materiał będzie zawierał walki bokserskie odbywające się w Polsce w 2021 roku podczas zawodów śląskiej ligi juniorów, kadetów i seniorów [98, 99, 100].

W celu przejścia do eksperymentów oraz nadzorowanego uczenia algorytmów zgromadzony materiał należały manualnie oznaczyć. Przed samym procesem oznaczania zebrane dane należy przetworzyć oraz dokonać procesu ich selekcji oraz transformacji. Następnie w celu oznaczania należy pozyskać licencjonowanych sędziów bokserskich, którzy z zastosowaniem wybranego narzędzia (po uprzednim przeszkoleniu) przez kolejne 6 miesięcy będą oznaczali zebrane nagrania. Wynikiem procesu oznaczania będzie baza danych zawierająca 312 774 oznaczonych klatek wideo.

Dopiero z tak przygotowaną bazą danych można będzie następnie rozpocząć dalsze eksperymenty oraz kolejne etapy procesu KDD. Cały proces budowania bazy danych wraz ze szczegółami jest opisany w sekcji 4.

Struktura rozprawy

Rozprawa składa się z 6 rozdziałów, z których rozdział 1 zawierać będzie podstawowe pojęcia i techniki związane z przetwarzaniem obrazu. Rozdział rozpocznie się od matematycznej definicji obrazu oraz jego reprezentacji cyfrowej, a także omówienia różnych modeli kolorów stosowanych w analizie obrazów. Następnie opisane zostaną podstawowe operacje przetwarzania

obrazu, wraz z ich matematycznymi reprezentacjami. Na koniec rozdziału omówione będzie zagadnienie przetwarzania wideo, które jest kluczowe dla dalszej części rozprawy.

Rozdział 2 będzie stanowił wprowadzenie do zagadnień związanych z uczeniem maszynowym, kluczowym obszarem informatyki, który obejmuje tworzenie algorytmów zdolnych do uczenia się z danych. W rozdziale omówione zostaną różne typy algorytmów uczenia maszynowego, w tym algorytmy nadzorowane i nienadzorowane, oraz ich zastosowania w praktycznych problemach, takich jak klasyfikacja i regresja. Szczegółowo opisane zostanie również pojęcie systemu informacyjnego w kontekście uczenia maszynowego, obejmujące zarówno dane wejściowe, jak i wyniki modeli. Dodatkowo, przedstawione zostaną etapy budowy modeli, techniki regularyzacji i zapobiegania nadmiernemu dopasowaniu, a także omówione zostaną metody oceny wydajności modeli.

W rozdziale 3 przedstawiona zostanie problematyka analizy zachowań zawodników w boksie olimpijskim, podkreślając zastosowanie technologii wizyjnych i uczenia maszynowego. Przetwarzanie obrazów znajduje szerokie zastosowanie w różnych dyscyplinach sportowych, umożliwiając automatyczne wydobywanie kluczowych informacji z materiałów wideo. W boksie olimpijskim analiza zachowań zawodników, takich jak rozpoznawanie ciosów i ocena strategii, staje się coraz bardziej istotna dla treningu i oceny technik, wymagając zaawansowanych narzędzi i zrozumienia specyfiki dyscypliny. W rozdziale zostaną omówione aktualne podejścia i wyzwania oraz znaczenie baz danych w badaniach nad przetwarzaniem obrazu w sporcie.

Rozdział 4 będzie zawierał szczegółowy opis znaczenia danych w algorytmach uczenia maszynowego, szczególnie w kontekście przetwarzania i klasyfikacji obrazów. Zostanie także omówione znaczenie wysokiej jakości oznaczonych danych dla efektywności modeli uczenia nadzorowanego oraz wyzwania związane z procesem zbierania i oznaczania danych. Zostanie przedstawiona specyfika zbioru danych użytego w tej rozprawie, metody jego pozyskania oraz wyzwania związane z dostępem do danych, w tym ograniczenia prawne i etyczne. Rozdział zawierać będzie również szczegółowy opis używanych narzędzi do oznaczania danych i proces organizacji pracy związanej z oznaczaniem danych w kontekście analizy ciosów w walkach bokserskich.

W rozdziale 5 zaprezentowany zostanie kompleksowy opis działań przeprowadzonych w celu realizacji głównego celu rozprawy, jakim jest analiza walk bokserskich. Rozdział zawiera opis etapów prac badawczych obejmujących wykrywanie bokserów na ringu, detekcję starć między zawodnikami oraz eliminację nieistotnych fragmentów wideo. Proces ten będzie kluczowy dla stworzenia precyzyjnie oznaczonej bazy danych, która posłuży do trenowania algorytmów uczenia maszynowego w sposób nadzorowany.

W rozdziale 6 zostanie skupiona uwaga na wyzwaniach związanych z klasyfikacją klatek wideo w kontekście boksu, gdzie kluczowe obszary są bardzo małe. Opisane zostaną problemy napotkane przy stosowaniu konwolucyjnych sieci neuronowych do klasyfikacji pojedynczych klatek oraz przedstawione zostanie autorskie podejście do segmentacji obrazu, które znacząco poprawia szybkość przetwarzania danych. Porównane zostaną różne techniki segmentacji, podkreślając ich wpływ na wydajność klasyfikacji oraz efektywność przetwarzania w czasie rzeczywistym. Wyniki eksperymentów wskażą na skuteczność proponowanych metod w poprawie jakości analizy bokserskiej, jednocześnie umożliwiając szybkie przetwarzanie danych, co ma kluczowe znaczenie dla dynamicznych sportów.

1. Przetwarzanie obrazu

W niniejszym rozdziale zostanie zawarty wstęp do tematyki obrazu oraz technik jego przetwarzania. Ponadto matematycznie zostanie zdefiniowany obraz oraz jego reprezentacja cyfrowa. Zostanie przedstawionych kilka modeli kolorów stosowanych do opisywania obrazów kolorowych. Ponadto zostaną zaprezentowane podstawowe operacje przetwarzania obrazu razem z ich matematycznymi reprezentacjami. Na końcu rozdziału zaprezentowane zostanie zagadnienie przetwarzania video, które jest przedmiotem dalszej części pracy.

1.1. Obraz

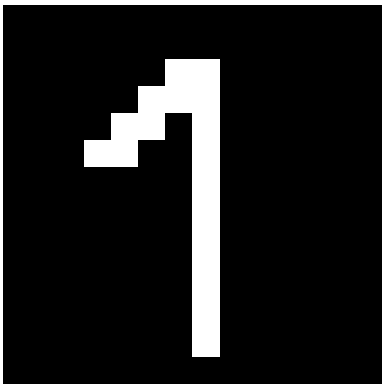
Obraz to reprezentacja wizualna, która jest odzwierciedleniem rzeczywistości, stworzona za pomocą różnych technik, takich jak rysowanie, fotografowanie, malowanie, czy też twórczość cyfrowa. Obraz może zawierać konkretne obiekty, osoby, sceny lub abstrakcyjne koncepcje [56, 72].

Obraz można zdefiniować jako dwuwymiarową funkcję $f(x, y)$, gdzie x i y są współrzędnymi przestrzennymi (płaszczyznowymi), a amplituda f w dowolnej parze współrzędnych (x, y) nazywana jest intensywnością lub poziomem szarości obrazu w tym punkcie. Gdy x, y i wartości intensywności f są skończonymi, dyskretnymi wielkościami, obraz nazywa się obrazem cyfrowym. Problem cyfrowego przetwarzania obrazu odnosi się do przetwarzania obrazów cyfrowych za pomocą komputera cyfrowego. Należy pamiętać, że obraz cyfrowy składa się ze skończonej liczby elementów, z których każdy ma określoną lokalizację i wartość. Elementy te nazywane są elementami obrazu, czyli pikselami. Piksel jest terminem najczęściej używanym do określenia elementów obrazu cyfrowego [82].

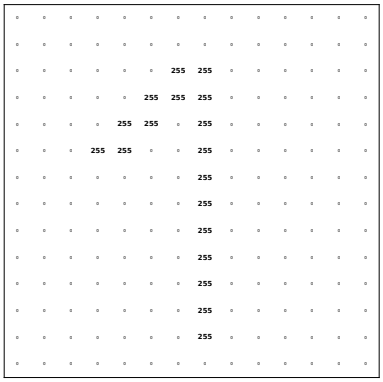
Obraz cyfrowy, $f(x, y)$, zawiera M wierszy i N kolumn, gdzie (x, y) są dyskretnymi współrzędnymi wskazującymi na pojedynczy element obrazu. Dla przejrzystości i wygody notacji używa się wartości całkowitych dla tych dyskretnych współrzędnych: $x = 0, 1, 2, ..., M - 1$ i $y = 0, 1, 2, ..., N - 1$. Tak więc, na przykład, wartość obrazu cyfrowego w punkcie początkowym to $f(0, 0)$, a jego wartość w kolejnych współrzędnych wzdłuż pierwszego wiersza to $f(0, 1), f(0, 2), ..., f(0, N - 1)$.

Reprezentacja obrazu na rysunku 1.1 oraz 1.6 jest najbardziej powszechna i pokazuje $f(x, y)$ tak, jak wyglądałaby na wyświetlaczu komputera lub na fotografii. Na rysunku 1.1 intensywność każdego punktu na wyświetlaczu jest proporcjonalna do wartości f w danym punkcie. Na tym rysunku (1.1) są tylko dwie rozmieszczone wartości intensywności. Jeśli intensywność jest znormalizowana do przedziału [0, 1], to każdy punkt na obrazie ma wartość 0 lub 1. Monitor lub drukarka konwertuje te dwie wartości odpowiednio na czerń lub biel.

Inna forma reprezentacji obrazu została przedstawiona na rysunku 1.2, jest to tablica (macierz) składająca się z wartości liczbowych $f(x, y)$. Jest to reprezentacja stosowana w przetwarzaniu komputerowym. W postaci równania zapisuje się reprezentację takiej tablicy



Rysunek 1.1: Odręcznie napisana cyfra 1 w skali szarości



Rysunek 1.2: Reprezentacja pikseli w skali szarości

numerycznej $M \times N$ jako:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & \cdots & f(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1,N-1) \end{bmatrix} \tag{1.1}$$

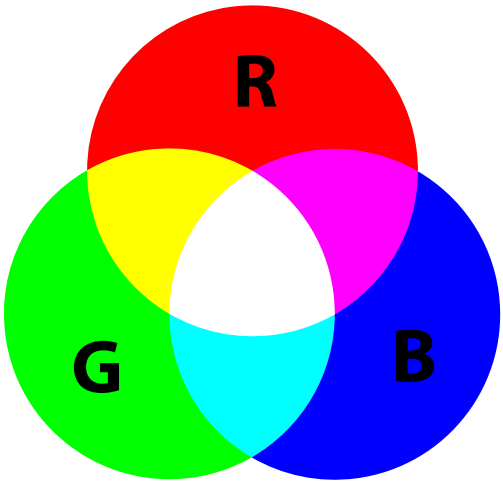
Prawa strona równania (1.1) to obraz cyfrowy reprezentowany jako tablica liczb rzeczywistych. Każdy element tej tablicy nazywany jest elementem obrazu lub pikselem. W całej pracy używane są terminy obraz i piksel do określenia obrazu cyfrowego i jego elementów. Konkretne piksele są wartościami tablicy w określonej parze współrzędnych. Obraz cyfrowy może być również reprezentowany w tradycyjnej formie macierzy:

$$A = \begin{bmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,N-1} \\ a_{1,0} & a_{1,1} & \cdots & a_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M-1,0} & a_{M-1,1} & \cdots & a_{M-1,N-1} \end{bmatrix} \tag{1.2}$$

gdzie $a_{ij} = f(i,j)$, więc równania (1.1) oraz (1.2) opisują identyczne macierze.

Środek obrazu cyfrowego $M \times N$ z początkiem w punkcie $(0, 0)$ i zakresem do $(M-1, N-1)$ uzyskuje się, dzieląc M i N przez 2 i zaokrąglając w dół do najbliższej liczby całkowitej. Operacja ta jest czasami oznaczana za pomocą operatora floor, $[\bullet]$. Dotyczy to zarówno parzystych, jak i nieparzystych wartości M i N. Na przykład środek obrazu o rozmiarze 1023 $\times$ 1024 znajduje się w punkcie (511, 512). Niektóre języki programowania (np. MATLAB) rozpoczynają indeksowanie od 1 zamiast od 0. W takim przypadku środek obrazu znajduje się w punkcie $(xc, yc) = (floor(M/2) + 1, floor(N/2) + 1)$.

W obrazie w skali szarości, wartość piksela reprezentuje poziom jasności, gdzie 0 (wartość minimalna) reprezentuje czarny, a 255 (wartość maksymalna) reprezentuje biały. Przykładem może być ręcznie napisana cyfra zapisana na rysunku 1.1, której siatka pikseli oraz ich wartości widnieje na rysunku 1.2. W kontekście obrazów kolorowych, modele kolorów odgrywają kluczową rolę, stanowiąc ramy do opisu i reprodukcji kolorów, z których można wyróżnić następujące modele [53, 78]:



Rysunek 1.3: Model kolorów RGB

- RGB (Red, Green, Blue): Model RGB jest powszechnie stosowany w urządzeniach stosujących emisję światła, takich jak monitory komputerowe, telewizory i ekrany smartfonów. RGB jest modelem adytywnym, w którym kolor generowany jest poprzez dodawanie intensywności składowych czerwonej (R), zielonej (G) i niebieskiej (B). Każdy piksel w obrazie RGB jest charakteryzowany przez trzy wartości, odpowiadające intensywności trzech kolorów składowych. W sposób graficzny model RGB został zaprezentowany na ryskunku 1.3.

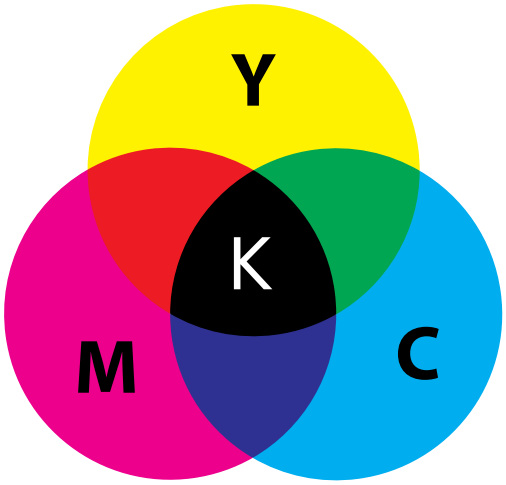
- CMYK (Cyan, Magenta, Yellow, Key/Black): Model CMYK jest najczęściej stosowany w technologii druku, ze szczególnym uwzględnieniem druku offsetowego. Jest to model subtraktywny, gdzie kolor tworzony jest przez odejmowanie światła za pomocą absorpcji. Cyjan (C), magenta (M) i żółty (Y) to barwy subtraktywne, a czarny (K) jest dodany do poprawy głębi i szczegółowości obrazu, jego graficzna reprezentacja została zaprezentowana na rysunku 1.4.

- HSV (Hue, Saturation, Value) / HSL (Hue, Saturation, Lightness): Modele HSV i HSL są często stosowane w aplikacjach do przetwarzania i analizy obrazów, oferując reprezentację kolorów, która jest intuicyjnie zrozumiała dla ludzkiego postrzegania. „Hue" odnosi się do odcienia koloru, „Saturation" do nasycenia, a „Value" lub „Lightness" odnosi się do jasności koloru. Model HSV został zaprezentowany na rysunku 1.5.

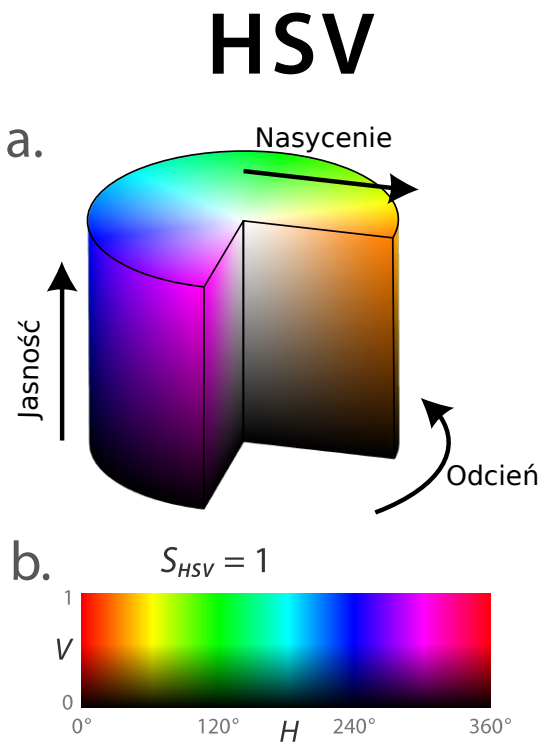
Przykładem obrazu kolorowego może być ręcznie napisana cyfra, która następnie została zapisana na zdjęciu o wymiarach 14x14 pikseli 1.6. Zdjęcie zostało zapisane w modelu RGB, a jego siatka z pikselami i wartościami dla każdego z kanałów została przedstawiona na zdjęciu 1.7. W pewnych zastosowaniach niezbędna jest konwersja obrazu zapisanego w modelu RGB na obraz w skali szarości, aby tego dokonać należy zastosować wzór (1.3).

$$f(x, y) = 0,299 \cdot R + 0,587 \cdot G + 0,114 \cdot B \tag{1.3}$$

gdzie $f(x, y)$ jest funkcją opisującą obraz, a R, G, B to kolejno wartości kanału koloru czerwonego, zielonego i niebieskiego.

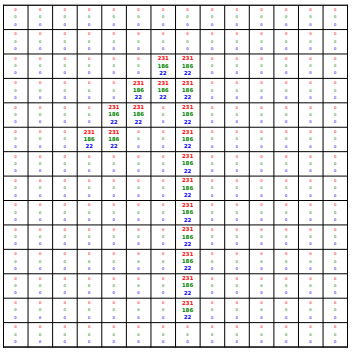


Rysunek 1.4: Model kolorów CMYK



Rysunek 1.5: Model kolorów HSV

Rysunek 1.6: Odręcznie napisana cyfra 1 w kolorze



Rysunek 1.7: Reprezentacja pikseli w skali RGB

Należy zauważyć, że każdy z tych modeli ma swoje unikalne zastosowania i ograniczenia, a wybór modelu kolorów zależy od konkretnego problemu. Dodatkowo, transformacje między różnymi modelami kolorów są często niezbędne w ramach różnych zastosowań przetwarzania obrazów.

Obrazy mogą być także trójwymiarowe, takie jak te używane w grafice komputerowej lub medycynie (np. tomografia komputerowa czy rezonans magnetyczny), gdzie wartość w każdym punkcie w przestrzeni reprezentuje na przykład gęstość tkanki [103]. Ponadto obrazy są fundamentalnym elementem w wielu dziedzinach nauki i techniki, takich jak grafika komputerowa, przetwarzanie obrazów, analiza obrazów, wizja komputerowa, medycyna, astronomia, i wiele innych [56, 72].

1.2. Przekształcenia na obrazach

Przetwarzanie obrazu to problem skupiający się na manipulacji i analizie obrazów cyfrowych za pomocą algorytmów. Jest to istotny aspekt wielu dziedzin, takich jak medycyna, robotyka, sztuczna inteligencja czy teledetekcja [78, 96, 105].

Jedną z podstawowych technik przetwarzania obrazu jest interpolacja, stosowana w zadaniach przetwarzania obrazu takich jak powiększanie, zmniejszanie, obracanie i geometryczna korekta obrazów. Interpolacja to proces stosowania znanych danych do szacowania wartości w nieznanych lokalizacjach. Dla przykładu zdjęcie o wymiarach 250x250 pikseli można powiększyć 2 krotnie do wymiarów 500x500 pikseli. W tym celu należy na nową siatkę pikseli 500x500 nałożyć oryginalne zdjęcie 250x250 pikseli, a następnie rozciągnąć do wymaganych rozmiarów. Takie przekształcenie wygeneruje nową siatkę pikseli o wymiarach 500x500 z pustymi elementami obrazu (pikselami), które następnie należy wypełnić.

Metod wypełniania pustych elementów jest wiele [82], jedna z nich polega na przypisywaniu wartości od najbliższego piksela dlatego nazywa się również interpolacją najbliższego sąsiada (ang. nearest neighbor interpolation). Podejście to jest proste, lecz ma tendencję do tworzenia niepożądanych artefaktów, takich jak zniekształcenia prostych krawędzi. Dokładniejszym podejściem jest interpolacja dwuliniowa (ang. bilinear interpolation), w której stosowanych jest czterech sąsiadów do estymacji wartości dla danego elementu. Niech (x, y) oznaczają

współrzędne elementu, któremu należy przypisać wartość oraz niech $v(x, y)$ oznacza wartość tego elementu. Dla interpolacji dwuliniowej, przypisywana wartość jest obliczana na podstawie równania (1.4).

$$v(x, y) = ax + by + cxy + d \tag{1.4}$$

gdzie cztery współczynniki a, b, c i d określają czterech najbliższych sąsiadów punktu (x, y). Interpolacja dwuliniowa daje znacznie lepsze wyniki niż interpolacja najbliższego sąsiada, przy niewielkim wzroście złożoności obliczeniowej [82].

Kolejnym poziomem złożoności jest interpolacja dwusześcienna (ang. bicubic interpolation), która na podstawie szesnastu najbliższych sąsiadów określa wartość przewidywanego punktu. Wartość danego punktu (x, y) jest uzyskiwana przy użyciu równania (1.5) [82].

$$v(x, y) = \sum_{i=0}^{3} \sum_{j=0}^{3} a_{ij} x^i y^j \tag{1.5}$$

Inną z przydatnych operacji w przetwarzaniu obrazu jest mierzenie dystansu pomiędzy pikselami np. a i b, których współrzędne następnie są określane jako (x, y), (u, v). Chcąc następnie zmierzyć odległość $d(a, b)$ pomiędzy pikselami a i b można zastosować miarę odległości Euklidesowej, która jest wyrażona wzorem (1.6).

$$d(a, b) = \sqrt{(x - u)^2 + (y - v)^2}, \tag{1.6}$$

Ważnym segmentem przetwarzania obrazu są również operacje arytmetyczne na obrazach. Zakładając, że istnieją dwa obrazy $f(x, y)$ oraz $g(x, y)$, na których można wykonać następujące operacje arytmetyczne:

$$s(x, y) = f(x, y) + g(x, y), \tag{1.7}$$

$$d(x, y) = f(x, y) - g(x, y), \tag{1.8}$$

$$p(x, y) = f(x, y) \times g(x, y), \tag{1.9}$$

$$v(x, y) = f(x, y) \div g(x, y). \tag{1.10}$$

Operacje arytmetyczne przedstawione na równaniach (1.7) (1.8) (1.9) (1.10) są operacjami elementarnymi, co oznacza, że są wykonywane między odpowiednimi parami pikseli f i g dla $x = 0, 1, 2, ..., M - 1$ i $y = 0, 1, 2, ..., N - 1$. Tak jak wcześniej, M i N są rozmiarami wierszy i kolumn przetwarzanych obrazów. Oczywiście s, d, p i v są również obrazami o rozmiarze $M \times N$. Warto zauważyć, że arytmetyka obrazów w sposób właśnie zdefiniowany obejmuje obrazy o tym samym rozmiarze.

Przykładem zastosowania operacji arytmetycznych może być redukcja szumów na obrazie, która stosuje technikę uśredniania obrazu (ang. image averaging) jednocześnie stosując operację dodawania obrazów do siebie (zgodnie ze wzorem (1.7)), popularną dziedziną nauki stosującą tą technikę jest astronomia. Innym przykładem może być również wykrywanie różnic na dwóch obrazach poprzez operację odjęcia obrazów od siebie (zgodnie ze wzorem (1.8)), taką technikę stosuje się podczas przetwarzania zdjęć satelitarnych w celu porównania zdjęć z różnych okresów i wykryciu zmian pomiędzy nimi.

1.3. Operacje morfologiczne

Operacje morfologiczne w przetwarzaniu obrazów odnoszą się do zestawu operacji matematycznych, które są stosowane na obrazie w celu manipulacji jego kształtem, strukturą i cechami. Do podstawowych operacji morfologicznych należą erozja i dylatacja obrazu, które kolejno zostały opisane wzorami (1.11) i (1.12). Erozja skraca granice obiektów pierwszoplanowych, podczas gdy dylatacja je rozszerza. Te techniki mogą być łączone w bardziej złożone operacje, takie jak otwarcie, które zostało opisane wzorem (1.13) stosujące na obrazie najpierw operację erozji, a następnie dylatacji. Istnieje również operacja zamknięcia stosująca najpierw dylatację, a następnie erozję. Techniki te są stosowane do usuwania szumów i wygładzania kształtów [30, 84].

$$A \ominus B = \{z \in E \mid B_z \subseteq A\} \tag{1.11}$$

gdzie A jest obrazem poddawanym erozji przez element strukturalny (kernel) B, E jest siatką liczb całkowitych, B_z jest przekształceniem B przez wektor z, więc $B_z = \{b + z \mid b \in B\}, \forall z \in E$.

$$A \oplus B = \{z \in E \mid (B^s)_z \cap A \neq \varnothing\} \tag{1.12}$$

gdzie A jest obrazem poddawanym dylatacji przez element strukturalny (kernel) B, E jest siatką liczb całkowitych, B^s oznacza symetryczność B, czyli $B^s = \{x \in E \mid -x \in B\}$.

$$opening(A, B) = dilation(erosion(A, B), B) \tag{1.13}$$

gdzie A jest obrazem poddawanym operacji otwarcia przez element strukturalny (kernel) B, *dilation* jest operacją dylatacji, którą opisuje wzór (1.12), *erosion* jest operacją erozji, którą opisuje wzór (1.11).

Głównym celem stosowania operacji morfologicznych jest przetwarzanie i poprawa obrazów poprzez usuwanie szumów, wypełnianie luk lub dziur, wygładzanie krawędzi i ekstrakcję istotnych cech. Te operacje pomagają w poprawie jakości obrazu i przygotowaniu obrazów do dalszej analizy [30, 84].

Poprzez zastosowanie operacji morfologicznych, nadmiarowe lub puste części obrazu mogą zostać przetworzone, nieistotne piksele mogą zostać usunięte, a luki mogą zostać wypełnione. Pomaga to w poprawie zadań rozpoznawania i analizy obrazów poprzez zwiększenie klarowności i dokładności obrazu [30, 84].

1.4. Odejmowanie tła

Odejmowanie tła to technika stosowana w przetwarzaniu obrazów i wizji komputerowej, szczególnie w zastosowaniach związanych z monitorowaniem wideo. Celem tej techniki jest identyfikacja ruchomych obiektów na obrazie poprzez odejmowanie tła, które jest obrazem sceny bez ruchomych obiektów [4, 5, 10, 15, 27, 50, 59, 85, 110].

W praktyce, model tła jest często tworzony przez obliczenie średniej lub mediany z serii obrazów zarejestrowanych w czasie, kiedy nie ma ruchomych obiektów. Następnie, każdy nowy obraz jest porównywany z tym modelem tła stosując wzór (1.14). Piksele, które różnią się od

modelu tła, są klasyfikowane jako należące do ruchomych obiektów i jednocześnie pozostają bez zmian, natomiast pozostałe piksele należące do tła przyjmują wartość 0 (czarny).

$$v(x,y) = f(x,y) \wedge mask(x,y) \tag{1.14}$$

gdzie $v(x,y)$ jest funkcją opisującą obraz, który zawiera piksele obrazu $f(x,y)$, dla których wartość na obrazie $mask(x,y)$ jest niezerowa, pozostałe piksele (dla których wartość na obrazie $mask(x,y)$ jest zerowa/negatywna) przyjmują wartość 0 (czarny).

Technika ta jest często stosowana w systemach monitoringu bezpieczeństwa, gdzie ważne do analizy są tylko ruchome obiekty, takie jak pojazdy lub przechodnie. Może być również używana w innych zastosowaniach, takich jak analiza ruchu, śledzenie obiektów, czy rozpoznawanie zachowań.

Istnieją również rozwiązania stosujące wiele kamer, gdzie każda kamera generuje swój własny model tła, a następnie te modele są łączone w celu stworzenia bardziej kompleksowego modelu tła. W zastosowaniach, takich jak monitorowanie wideo, systemy z wieloma kamerami mogą zapewnić większe pokrycie obszaru i lepszą detekcję obiektów, ponieważ różne kamery mogą rejestrować różne części sceny z różnych perspektyw. Jednakże, implementacja takiego systemu może być skomplikowana, ponieważ wymaga kalibracji i synchronizacji między kamerami, a także specjalnych algorytmów do łączenia informacji z różnych kamer [60].

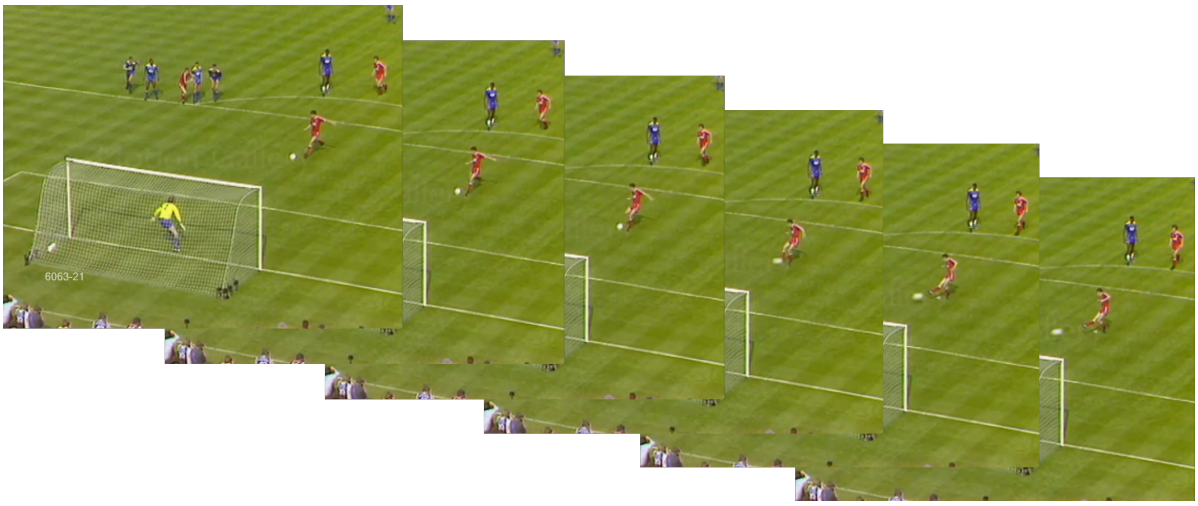
1.5. Odejmowanie tła dynamicznego

Osobną dziedziną nauki jest usuwanie tła dynamicznego (ang. dynamic background removal, DBR), jest to jeden z kluczowych i zarazem trudnych problemów w dziedzinie przetwarzania obrazu i analizy wideo. Środowisko dynamiczne, które obejmuje zmienne tło, takie jak fale wody, poruszające się liście, ruchome tłumy, zmienne warunki oświetleniowe, stanowią wyzwanie dla tradycyjnych technik odejmowania tła [11, 26, 109, 128].

Tradycyjne metody takie jak modelowanie tła za pomocą mieszanych modeli Gaussa (ang. gaussian mixture models, GMM), mogą okazać się niewystarczające, gdyż nie radzą sobie dobrze ze zmiennym tłem. GMM, stosując kombinację kilku rozkładów Gaussa do modelowania pikseli tła, mogą być skuteczne w środowiskach statycznych, ale napotykają trudności w sytuacjach dynamicznych, gdzie tło szybko się zmienia.

Metody stosujące uczenie maszynowe, w tym głębokie sieci neuronowe (ang. deep neural networks, DNN), wykazują lepsze wyniki [50, 106]. Metody te mogą uczyć się złożonych wzorców zmian w dynamicznym tle i lepiej radzić sobie z takimi zmianami w czasie rzeczywistym. Przykładowo, sieci konwolucyjne (ang. convolutional neural networks, CNN) mogą być stosowane do nauki lokalnych cech w obrazie, podczas gdy rekurencyjne sieci neuronowe (ang. recurrent neural networks, RNN) mogą bazować na informacjach z poprzednich klatek do lepszego przewidywania tła w przyszłości [80].

Jednak metody oparte na uczeniu maszynowym również napotykają wyzwania. Wymagają one dużych zbiorów danych treningowych i mocy obliczeniowej, a ponadto często napotykają na problemy takie jak przetrenowanie (ang. overfitting). Wyniki mogą też być trudne do zinterpretowania, co utrudnia ich użycie w aplikacjach wymagających wyjaśnialności, takich jak te stosowane w medycynie lub prawie.



Rysunek 1.8: Sekwencja klatek-obrazów nagrania wideo

Istnieje wiele aktywnych obszarów badań w celu rozwiązania tych problemów i poprawy efektywności technik DBR w środowiskach dynamicznych, takich jak rozwijanie nowych architektur sieci neuronowych, badanie różnych technik augmentacji danych, czy też badanie sposobów integracji informacji z różnych źródeł (np. z różnych kamer czy różnych ram czasowych).

1.6. Przetwarzanie wideo

W nagraniach wideo obraz odgrywa kluczową rolę dostarczając wizualne informacje, które są istotne dla percepcji i interpretacji treści przez odbiorcę. W kontekście przetwarzania wideo, obraz jest podstawowym elementem, na którym opierają się wszystkie analizy i manipulacje. Nagranie wideo składa się z sekwencji pojedynczych obrazów wyświetlanych w określonym tempie, co tworzy iluzję ciągłego ruchu dla ludzkiego oka. Każda klatka obrazu w nagraniu wideo jest obrazem statycznym, a zmiana klatek tworzy efekt ruchu [86, 112]. Taką sekwencję obrazów, które następnie składają się w wideo zawiera rysunek 1.8.

Przetwarzanie wideo, podobnie jak przetwarzanie obrazów statycznych obejmuje wiele technik, takich jak detekcja obiektów, śledzenie ruchu, segmentacja czy filtracja. Wideo jest jednak bardziej skomplikowane do analizy, ponieważ zawiera dodatkowy wymiar - czas.

Jednym z kluczowych zadań w przetwarzaniu wideo jest analiza ruchu, która polega na rozpoznawaniu i śledzeniu ruchu obiektów w sekwencji wideo. Analiza ruchu może być podzielona na dwie główne kategorie:

- detekcja ruchu - celem detekcji ruchu jest identyfikacja pikseli na obrazie, które uległy zmianie między dwoma kolejnymi klatkami lub przez pewien okres czasu. Prostym podejściem jest odejmowanie jednej klatki od drugiej i zastosowanie progowania do wyniku, aby zidentyfikować obszary, które uległy zmianie. Bardziej zaawansowane techniki mogą uwzględniać szumy i fluktuacje intensywności, które są nieistotne dla ruchu.

- Śledzenie ruchu - po zidentyfikowaniu obiektów w ruchu, następnym krokiem jest śledzenie ich pozycji i ruchu w czasie. To jest zazwyczaj trudniejsze zadanie, ponieważ obiekty mogą się poruszać w różny sposób, zmieniać swój kształt i wygląd lub podczas przemieszczania być przysłaniane przez inne obiekty. Wiele technik zostało opracowanych do śledzenia ruchu, od prostych metod opartych na dopasowaniu wzorców, do zaawansowanych technik opartych na filtrach Kalmana i cząstkach, aż po metody oparte na głębokim uczeniu.

Analiza ruchu jest podstawą w rozpoznawaniu zachowań ludzkich (ang. human action recognition, HAR). HAR to obszar badań w zakresie sztucznej inteligencji (ang. artificial intelligence, AI) i przetwarzania wideo, który koncentruje się na identyfikacji i klasyfikacji różnych działań wykonywanych przez ludzi na nagraniach wideo. Przykładowe akcje, które mogą być rozpoznawane to bieganie, skakanie, czytanie, czy też te rejestrowane przez kamery bezpieczeństwa takie jak kradzież, bójka czy włamanie. Zależy to przede wszystkim od środowiska w jakim system do rozpoznawania zachowań jest uruchomiony i jakie jest jego przeznaczenie [2, 71, 119].

HAR jest trudnym zadaniem z powodu dużej zmienności w działaniach ludzkich. Na przykład, to samo działanie może być wykonane w różny sposób przez różne osoby, w różnym tempie, z różnymi poziomami energii itp. Ponadto, różne akcje mogą wyglądać podobnie, zwłaszcza gdy są obserwowane z różnych perspektyw lub są częściowo zasłonięte przez inne obiekty [14, 76].

Techniki stosowane do rozpoznawania działań ludzkich zwykle polegają na uczeniu maszynowym i uczeniu głębokim. Na przykład, konwolucyjne sieci neuronowe (ang. convolutional neural networks, CNN) mogą być używane do ekstrakcji cech z obrazów wideo, a rekurencyjne sieci neuronowe (ang. recurrent neural networks, RNN) lub sieci neuronowe o długiej pamięci krótkotrwałej (ang. long short rerm memory, LSTM) mogą być używane do modelowania sekwencji czasowych tych cech.

2. Uczenie maszynowe

Uczenie maszynowe (ang. machine learning, ML) to naukowe badanie algorytmów obliczeniowych, które są zaprojektowane do wykonywania określonego zadania bez wyraźnego programowania. Algorytmy te zostały zaprojektowane w celu naśladowania ludzkiej inteligencji i rozwiązywania zadań, takich jak klasyfikacja, regresja czy grupowanie. Algorytmy uczenia maszynowego można podzielić na nadzorowane i nienadzorowane. Rodzaj uczenia się algorytmu zależy od struktury danych w konkretnym zadaniu uczenia maszynowego [73].

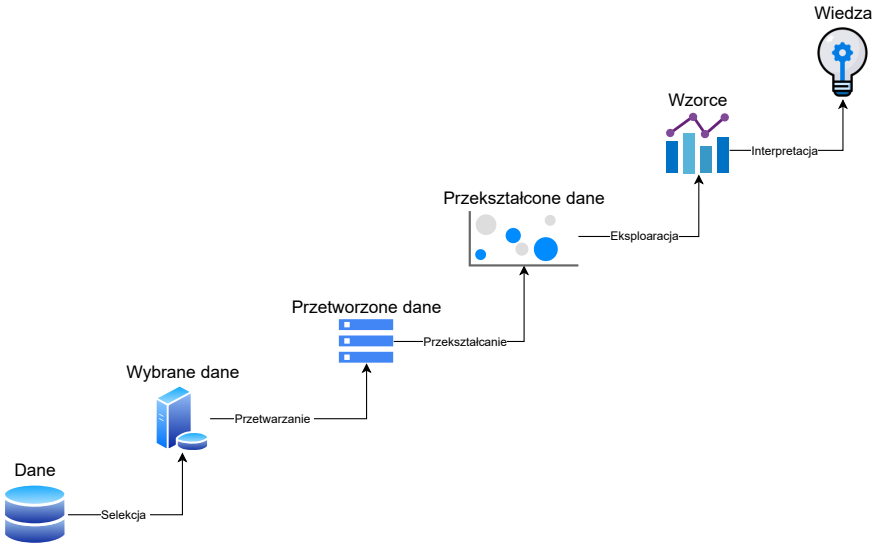
Uczenie nadzorowane jest stosowane do uczenia się na danych, w których podano pożądane dane wyjściowe, takie jak na przykład wykrywanie raka na podstawie skanów CT [103], gdzie każdy obraz (obserwacja) zawiera informacje (etykietę) o tym, czy pochodzi od zdrowego czy chorego na raka pacjenta. Tymczasem uczenie bez nadzoru opiera się na danych, w których etykiety nie są znane, a algorytmy nie mają pożądanych danych wyjściowych, takich jak grupowanie wina na podstawie 13 cech opisujących skład chemiczny [93].

Niniejszy rozdział zawierać będzie wstęp do problemu odkrywania wiedzy z danych oraz przegląd algorytmów i technik uczenia maszynowego, wraz z teoretycznymi podstawami, praktycznymi aspektami implementacji oraz przeglądem technik oceny ich wydajności. Szczegółowo zostaną opisane etapy budowy modeli, techniki regularyzacji i zapobiegania nadmiernemu dopasowaniu z podsumowaniem wniosków istotnych dla dalszych rozdziałów rozprawy.

2.1. Odkrywanie wiedzy z danych

Problem odkrywania wiedzy z danych (ang. knowledge discovery in databases, KDD) zajmuje się opracowywaniem metod i technik wydobywania informacji z danych. Podstawowym problemem rozwiązywanym przez proces KDD jest mapowanie danych niskiego poziomu (które są zazwyczaj zbyt obszerne, aby się łatwo z nimi zaznajomić i zrozumieć) na inne formy, które mogą być bardziej zwięzłe (na przykład krótki raport), bardziej abstrakcyjne (na przykład skrótowy opis danych) lub bardziej użyteczne (na przykład model predykcyjny do szacowania wartości przyszłych przypadków) [23]. Według [22] KDD to nietrywialny proces identyfikowania prawidłowych, nowatorskich, potencjalnie użytecznych i ostatecznie zrozumiałych wzorców w danych.

Tradycyjna metoda przekształcania danych w wiedzę opiera się na ręcznej analizie i interpretacji. Na przykład w branży opieki zdrowotnej specjaliści okresowo analizują bieżące trendy i zmiany w danych dotyczących opieki zdrowotnej na przykład co kwartał. Następnie specjaliści dostarczają raport zawierający szczegóły analizy do sponsorującej organizacji opieki zdrowotnej; raport ten staje się podstawą do podejmowania przyszłych decyzji i planowania zarządzania opieką zdrowotną. W zupełnie innej dziedzinie geolodzy planetarni manualnie przeglądają obrazy planet i asteroid, a następnie starannie lokalizują i kategoryzują takie obiekty geologiczne, jak kratery uderzeniowe. Dlatego też niezależnie od dziedziny klasyczne podejście do analizy danych opiera się zasadniczo na tym, że jeden lub więcej analityków zapoznaje się



Rysunek 2.1: Proces odkrywania wiedzy z danych (KDD)

z danymi i służy jako łącznik między danymi a użytkownikami i produktami [22, 23].

W aspekcie tych i innych zastosowań, metoda manualnej analizy zbiorów danych charakteryzuje się niską efektywnością, znacznymi kosztami oraz wysokim stopniem subiektywności. W obliczu rosnącej eksponencjalnie ilości danych, manualny sposób analizy danych staje się w wielu dziedzinach niepraktyczny lub całkiem niemożliwy [22, 23].

KDD to kompleksowy proces wydobywania przydatnej wiedzy z danych, przy czym eksploracja danych odnosi się do jednego z jego kluczowych etapów. Eksploracja danych polega na stosowaniu określonych algorytmów do identyfikowania wzorców w danych. Proces KDD jest działaniem interdyscyplinarnym, wymagającym technik wykraczających poza granice jednej specyficznej dziedziny, na przykład eksploracji danych czy uczenia maszynowego. W tym kontekście, inne obszary sztucznej inteligencji (poza uczeniem maszynowym) mają istotny potencjał do wniesienia wkładu w rozwój KDD. Istotnym aspektem KDD jest skupienie na odkrywaniu wzorców, które są zrozumiałe i mogą być interpretowane jako użyteczna lub ciekawa wiedza. Na przykład, sieci neuronowe, mimo że są efektywnym narzędziem modelowania, mogą być trudniejsze do zrozumienia niż drzewa decyzyjne. KDD zwraca także uwagę na skalowalność i odporność algorytmów modelowania w przypadku dużych i zaszumionych zbiorów danych [22, 23].

Proces KDD obejmuje bazę danych oraz wszystkie niezbędne etapy, takie jak selekcja danych, wstępne przetwarzanie, tworzenie podpróbek i przekształcenia danych. Następnie stosuje się metody eksploracji danych, czyli różnorodne algorytmy, w celu wydobycia z bazy danych określonych wzorców. Kolejnym etapem jest ocena wyników eksploracji danych, aby zidentyfikować powstałe wzorce, które można uznać za istotną wiedzę. Komponent eksploracji danych w procesie KDD koncentruje się na algorytmicznych sposobach wyodrębniania i obliczania wzorców z danych. Cały proces KDD zawiera rysunek 2.1, który obejmuje także ocenę i potencjalną interpretację odkrytych wzorców, aby ustalić, które z nich mogą być traktowane jako nowa wiedza [22, 23].

Proces KDD jest interaktywny i iteracyjny, obejmujący wiele kroków z wieloma decyzjami

podejmowanymi przez użytkownika. Autorzy [8] przedstawiają praktyczne spojrzenie na proces KDD, podkreślając jego interaktywny charakter. Składa się on z następujących elementów:

- zrozumienie specyfiki dziedziny, w której ma być wykorzystany proces KDD oraz określenie celu procesu z punktu widzenia interesanta.
- Stworzenie zestawu danych, na którym odbywać się będzie proces KDD - wybranie zestawu danych lub skupienie się na podzbiorze zmiennych lub próbek istniejącej bazy danych, na których ma zostać przeprowadzone odkrywanie.
- Czyszczenie i wstępne przetwarzanie danych. Podstawowe operacje obejmują usuwanie szumu oraz obserwacji odstających, zbieranie informacji niezbędnych do modelowania, podejmowanie decyzji dotyczących strategii obsługi brakujących pól w danych oraz uwzględnianie informacji o sekwencji czasowej i znanych zmianach.
- Redukcja wymiarowości przetwarzanych danych - znalezienie użytecznych cech do reprezentowania danych w zależności od celu zadania. Dzięki redukcji wymiarowości lub metodom transformacji można zmniejszyć efektywną liczbę rozważanych zmiennych lub znaleźć niezmienne reprezentacje danych.
- Dopasowanie celów procesu KDD (krok pierwszy) do konkretnej metody eksploracji danych, na przykład klasyfikacja, regresja, grupowanie [22].
- Analiza eksploracyjna oraz wybór modelu i hipotezy: wybór algorytmu/algorytmów eksploracji danych i wybór metody/metod do wyszukiwania wzorców danych. Proces ten obejmuje podjęcie decyzji, które modele i parametry mogą być odpowiednie (na przykład modele danych kategorycznych różnią się od modeli wektorów w liczbach rzeczywistych) oraz dopasowanie konkretnej metody eksploracji danych do ogólnych kryteriów procesu KDD (na przykład interesant może być bardziej zainteresowany zrozumieniem modelu niż jego możliwościami predykcyjnymi).
- Eksploracja danych: wyszukiwanie interesujących wzorców w określonej formie reprezentacji lub zestawie takich reprezentacji, w tym reguł klasyfikacji lub drzew, regresji i grupowania. Interesant może znacząco wspomóc metodę eksploracji danych poprzez prawidłowe wykonanie poprzednich kroków.
- Interpretacja wydobytych wzorców, ewentualnie powrót do któregokolwiek z poprzednich kroków w celu dalszej iteracji. Ten etap może również obejmować wizualizację wyodrębnionych wzorców i modeli lub wizualizację danych z uwzględnieniem wyodrębnionych modeli.
- Działanie na odkrytej wiedzy: bezpośrednie wykorzystanie wiedzy, włączenie wiedzy do innego systemu w celu dalszego działania lub po prostu udokumentowanie jej i zgłoszenie faktu interesantom. Proces ten obejmuje również sprawdzanie i rozwiązywanie potencjalnych konfliktów z wcześniej uznaną wiedzą.

Element eksploracji danych w ramach procesu KDD jest zazwyczaj charakteryzowany przez stosowanie wielokrotnych, iteracyjnych metod analizy danych. Specyfikacja celów procesu

KDD jest ściśle powiązana z przewidywanym zastosowaniem systemu eksploracyjnego. W tym kontekście możemy wyróżnić dwa główne typy celów: weryfikację i odkrywanie. W przypadku weryfikacji system eksploracji danych koncentruje się na testowaniu i potwierdzaniu hipotez sformułowanych przez interesanta. Proces ten polega na sprawdzeniu, czy istniejące przypuszczenia odnoszące się do zestawu danych znajdują empiryczne potwierdzenie w analizowanych danych. Natomiast proces odkrywania charakteryzuje się tym, że system analizy danych działa w sposób autonomiczny, mając na celu identyfikację nowych, wcześniej nieznanych wzorców i zależności w danych. Ta forma eksploracji nie opiera się na wcześniej sformułowanych hipotezach, a raczej pozwala systemowi na samodzielne generowanie nowych tez i koncepcji na podstawie analizy.

Eksploracja danych to proces, który obejmuje adaptację modeli analitycznych do istniejących zbiorów danych oraz identyfikację wzorców opartych na tych danych. Znaczącą rolę w tym kontekście odgrywa dopasowanie modeli, które przekłada się na wypracowanie wiedzy opartej na analizie danych. Kluczowe jest tutaj zrozumienie, że użyteczność i relewancja tych modeli w kontekście wiedzy wynikają z interaktywnego procesu KDD, który często wymaga subiektywnej oceny ze strony człowieka. W kontekście dopasowywania modeli do danych wyróżnia się dwa podstawowe formalizmy matematyczne [8, 22, 23]:

- formalizm statystyczny - to podejście akceptuje istnienie elementu losowości i niepewności w modelach. Pozwala na uwzględnienie niedeterministycznych efektów, co jest szczególnie użyteczne w sytuacjach, gdy dane zawierają elementy losowości lub są niepełne.

- Formalizm logiczny - w przeciwieństwie do podejścia statystycznego, formalizm logiczny charakteryzuje się determinizmem. Oznacza to, że model generuje wyniki w sposób jednoznaczny i przewidywalny, bazując na zdefiniowanych zasadach logicznych.

W praktycznych zastosowaniach eksploracji danych dominuje podejście statystyczne, które jest szczególnie adaptowalne do rzeczywistych scenariuszy, gdzie często spotyka się niepewności oraz zmienność w generowanych danych. Ten formalizm jest zatem preferowany ze względu na jego zdolność do efektywnego radzenia sobie z typowymi wyzwaniami i niepewnościami, które pojawiają się w procesie analizy danych.

Szereg różnych algorytmów może być często oszałamiający zarówno dla nowicjuszy, jak i doświadczonych analityków danych. Należy podkreślić, że spośród wielu metod eksploracji danych opracowanych w literaturze, tak naprawdę istnieje tylko kilka podstawowych technik. Rzeczywista podstawowa reprezentacja modelu używana przez konkretną metodę zazwyczaj pochodzi z kompozycji niewielkiej liczby dobrze znanych opcji: wielomianów, splajnów, funkcji jądra, funkcji bazowych, funkcji progowych, logicznych itp. W związku z tym algorytmy różnią się przede wszystkim kryterium dobrego dopasowania stosowanym do oceny dopasowania modelu lub metodą wyszukiwania stosowaną do znalezienia dobrego dopasowania [8, 22, 23].

Większość metod eksploracji danych opiera się na wypróbowanych i przetestowanych technikach uczenia maszynowego, rozpoznawania wzorców i statystyki takich jak: klasyfikacji, grupowaniu, regresji i tak dalej. Opis przykładowych algorytmów stosowanych w procesie eksploracji danych został opisany w dalszej części niniejszego rozdziału.

2.2. Problem klasyfikacji

Klasyfikacja to powszechnie stosowana technika w uczeniu maszynowym, mająca szerokie zastosowanie w różnych dziedzinach nauki. Polega ona na przyporządkowywaniu każdego przypadku w zbiorze danych do jednej z predefiniowanych klas. Jest to kluczowa funkcja w procesie eksploracji danych, która umożliwia przypisanie elementów do odpowiednich kategorii lub klas. Według [48] celem klasyfikacji jest dokładne przewidzenie klasy docelowej dla każdego przypadku w danych. Według innej definicji klasyfikacja jest procesem uczenia się funkcji, która mapuje (klasyfikuje) element danych do jednej z kilku predefiniowanych klas [117].

Przykładem problemu klasyfikacji może być klasyfikator do identyfikacji ryzyka kredytowego (np.: niskiego, średniego i wysokiego) dla wnioskujących o kredyt. Zadanie klasyfikacji rozpoczyna się od pozyskania zbioru danych, który zawiera obserwacje oraz przypisane klasy (ryzyka zdefiniowane przez człowieka). Na przykład model klasyfikacji, który przewiduje ryzyko kredytowe, może zostać opracowany na podstawie danych zaobserwowanych dla wielu wnioskodawców kredytowych w pewnym historycznym okresie czasu. Model klasyfikacji przewiduje przynależność do klasy danego wnioskodawcy na podstawie cech klienta, w tym przypadku może to być np. historia zatrudnienia, forma zatrudnienia, wysokość wynagrodzenia, posiadane nieruchomości itp. Model klasyfikacyjny przewiduje wartości dyskretne.

Najprostszym typem problemu klasyfikacji jest klasyfikacja binarna. W klasyfikacji binarnej atrybut docelowy ma tylko dwie możliwe wartości: na przykład niski lub wysoki stopień ryzyka przyznania kredytu. Klasyfikacje wieloklasowe mają więcej niż dwie wartości, tak jak w przykładzie wcześniejszym, stopień ryzyka może być: niski, średni lub wysoki. W procesie budowania/uczenia modelu, algorytm klasyfikacji znajduje relacje między wartościami cech, a wartościami celu. Różne algorytmy klasyfikacji stosują różne techniki znajdowania relacji. Relacje te są podsumowywane w modelu, który można następnie zastosować do innego zestawu danych, w którym przypisania klas nie są wcześniej znane. Klasyfikacja ma wiele zastosowań w segmentacji klientów, marketingu, analizie kredytowej, modelowaniu reakcji biomedycznych i lekowych czy też klasyfikacji obrazów [48]. Przykładem klasyfikacji może być rysunek 2.2, który przedstawia problem klasyfikacji binarnej spam/nie spam dla wiadomości email.

W klasyfikacji nadzorowanej otrzymuje się zbiór próbek (zwany także zbiorem uczącym). Zbiór ten składa się z n obserwacji (zwanych również obiektami lub próbkami):

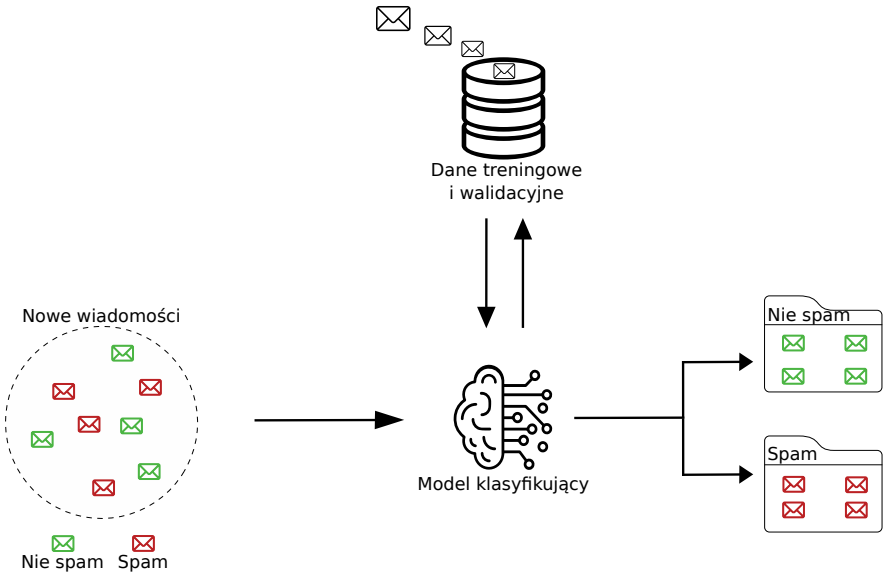
$$X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}, \tag{2.1}$$

Każda z obserwacji x_i jest opisana przez m atrybutów (zwanych również cechami)

$$a_1, a_2, \ldots, a_m, \tag{2.2}$$

z $a_j \in A_j$, $j = 1, \ldots, m$, gdzie A_j oznacza domenę j-tego atrybutu. W ten sposób cechy $a_1, a_2, \ldots, a_m$ tworzą przestrzeń cech $A_1 \times A_2 \times A_m$.

Wartości tych atrybutów mogą być ilościowe (na przykład wartość rynku akcji) lub kategoryczne (na przykład ryzyko kredytowe: „niskie” lub „wysokie”). Każda obserwacja należy do jednej z C różnych (skończenie wielu) i znanych klas decyzyjnych. Dlatego każda obserwacja może być reprezentowana jako:



Rysunek 2.2: Przykład problemu klasyfikacji wiadomości email

$$x_i = (\vec{V_i}, c_i), v_i^j \in A_j, c_i \in \{1, \ldots, C\}, \tag{2.3}$$

gdzie $\vec{V_i} = [v_i^1, \ldots, v_i^m]$ jest wektorem w m-wymiarowej przestrzeni cech, v_i^j jest wartością atrybutu a_j dla obserwacji (obiektu) x_i, a c_i jest etykietą klasy (zwaną również klasą decyzyjną) tej obserwacji (obiektu x_i).

W związku z tym X można przedstawić jako:

$$X : \{(\vec{V_i}, c_i)\}_{i=1}^n. \tag{2.4}$$

W oparciu o powyższe definicje, problem klasyfikacji można zdefiniować jako określenie sposobu przypisania obiektu do klasy, wiedząc, że istnieje C różnych klas decyzyjnych i że każdy obiekt należy do jednej z nich. Algorytm uczący $\mathcal{L}$ jest najpierw trenowany na zbiorze wstępnie sklasyfikowanych przykładów X. W klasyfikacji każdy c_i przyjmuje jedną z C wartości nominalnych. X składa się z niezależnie i identycznie rozłożonych próbek uzyskanych zgodnie z ustalonym, ale nieznanym, wspólnym rozkładem prawdopodobieństwa κ_{c_i} w przestrzeni cech w każdej klasie.

Celem klasyfikacji jest stworzenie klasyfikatora, który może być wykorzystany do podzielenia zbioru obiektów na odrębne klasy (klasyfikacja obiektów), a ponadto do oceny przeprowadzonej klasyfikacji. Można więc powiedzieć, że w tym procesie proponowana jest hipoteza h, która najlepiej przybliża funkcję oceny F (w odniesieniu do wybranej miary jakości klasyfikacji, na przykład: dokładności lub precyzji). Oznacza to, że hipoteza h minimalizuje funkcję straty (tj. stratę zero-jedynkową) w przestrzeni wektorów cech i klas $\vec{V} \times C$ w oparciu o rozkład κ_{c_i}.

Klasyfikacja rozpoczyna się, gdy algorytm uczący $\mathcal{L}$ otrzymuje jako dane wejściowe zbiór uczący X i przeprowadza wyszukiwanie w przestrzeni hipotez $H_\mathcal{L}$, która przybliża prawdziwą funkcję F. Dokładniej, algorytm uczący jest odwzorowaniem $\mathcal{L} : SH_\mathcal{L}$, gdzie S jest przestrzenią wszystkich zbiorów uczących o rozmiarze n, która odwzorowuje zbiór uczący na hipotezę.

Wybrana hipoteza h może być następnie wykorzystana do przewidywania klasy niewidocznych przykładów [33].

Należy wspomnieć, że klasyfikacja, w swoim klasycznym podejściu, różni się od tak zwanego uczenia ze wzmocnieniem. W przypadku uczenia ze wzmocnienie algorytm uczenia się zawiera pewne procedury sprzężenia zwrotnego, które bezpośrednio informują algorytm o osiągniętej jakości. Ta metoda uczenia się umożliwia znajdowanie rozwiązania bez konieczności korzystania z wcześniej zdobytej wiedzy. W klasycznym podejściu klasyfikatory są generowane na podstawie zdobytej wiedzy (zbioru treningowego) bez żadnych dodatkowych informacji zwrotnych.

Zbiór treningowym można nazwać również tabelą decyzyjną, która jest prostą formą prezentacji pozyskanych obserwacji i informacji, które zostaną zastosowane w procesie uczenia algorytmu. Zapisane informacje nie determinują sposobu przetwarzania decyzji, odbiorcy decyzji oraz danych wejściowych i wyjściowych. Tabele decyzyjne służą głównie do przechowywania danych, a także są wykorzystywane do dalszej weryfikacji jakości klasyfikatora. Można więc powiedzieć, że tabela decyzyjna, zapisana jako uporządkowana para (równanie (2.5)), reprezentuje problem, dla którego budowany jest klasyfikator [57].

Przykładową tabelę decyzyjną można również przedstawić w postaci tabeli 2.1, w której podano zestaw atrybutów warunkowych i atrybut decyzyjny. Obserwacje są przedstawione jako obiekty $x_1 \dots x_6$.

$$(X, A \cup \{c\}), \tag{2.5}$$

gdzie X to zbiór obiektów, a A to zbiór atrybutów, w tym atrybut decyzyjny — c.

Tabela 2.1: Tabela decyzyjna

	Atrybuty				Atrybut decyzyjny
	a_1	a_2	a_3	a_4	c
x_1	1	0	0	1	1
x_2	1	0	1	1	0
x_3	0	0	0	1	1
x_4	0	0	1	0	0
x_5	1	1	0	0	0
x_6	0	1	0	0	1

2.3. Problem regresji

Regresja oraz grupowanie to dwa fundamentalne podejścia w dziedzinie uczenia maszynowego, które mają kluczowe znaczenie dla analizy i interpretacji danych. Techniki regresji oraz grupowania stosują różne metody do wydobywania wiedzy z danych, jednak obie odgrywają zasadniczą rolę w zrozumieniu zależności między zmiennymi oraz w identyfikacji struktur w zbiorach danych. Regresja, będąca techniką uczenia nadzorowanego, skupia się na przewidywaniu wartości ciągłej zmiennej zależnej na podstawie jednej lub więcej zmiennych niezależnych, umożliwiając prognozowanie oraz zidentyfikowanie relacji przyczynowo-skutkowych. Z kolei grupowanie to kluczowy element uczenia nienadzorowanego, który dąży do odkrycia

naturalnych podziałów w danych poprzez organizowanie obserwacji w klastry, co pozwala na identyfikację wzorców i struktur w danych bez wcześniej zdefiniowanych etykiet.

Regresja jest kluczową techniką w statystyce i uczeniu maszynowym, która pozwala na modelowanie i analizę związków między zmiennymi. Jest to proces statystyczny do estymowania relacji między zmienną zależną (Y) a jedną lub więcej zmiennymi niezależnymi (X). Celem regresji jest zrozumienie, jak wartości zmiennych niezależnych wpływają na zmienną zależną, co umożliwia przewidywanie lub prognozowanie nowych wartości.

2.3.1. Regresja liniowa

Regresja liniowa jest najprostszym i najczęściej stosowanym typem modelu regresji, który zakłada liniową relację między zmienną zależną, a zmiennymi niezależnymi. Model regresji liniowej można zdefiniować następująco:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{2.6}$$

gdzie:

- Y jest zmienną zależną,
- $X_1, X_2, \ldots, X_n$ są zmiennymi niezależnymi,
- β_0 jest wyrazem wolnym (przecięcie z osią y),
- $\beta_1, \beta_2, \ldots, \beta_n$ są współczynnikami kierunkowymi (pokazującymi wpływ każdej zmiennej niezależnej na Y),
- ϵ jest terminem błędu, reprezentującym różnicę między wartościami obserwowanymi a modelowanymi.

Regresja liniowa jest szeroko stosowana ze względu na swoją prostotę i skuteczność w wielu scenariuszach. Pozwala na łatwą interpretację współczynników modelu, gdzie każdy współczynnik β_i wskazuje, o ile zmieni się wartość zmiennej zależnej Y przy zmianie wartości odpowiedniej zmiennej niezależnej X_i o jednostkę, przy założeniu stałości pozostałych zmiennych [24, 39].

2.3.2. Regresja logistyczna

Regresja logistyczna jest używana, gdy zmienna zależna jest kategoryczna, na przykład w przypadkach, gdzie wynik jest typu tak/nie lub 0/1. Jest to forma regresji, która pozwala modelować prawdopodobieństwo przynależności do określonej kategorii. Model regresji logistycznej jest opisany równaniem:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{2.7}$$

gdzie p jest prawdopodobieństwem przynależności do jednej z kategorii. Funkcja logistyczna, czyli funkcja logit, transformuje prawdopodobieństwo p tak, aby jego zakres był w przedziale

$(-\infty, +\infty)$, co umożliwia stosowanie liniowych technik regresyjnych do modelowania danych kategorycznych.

Regresja, zarówno liniowa jak i logistyczna, odgrywa kluczową rolę w analizie danych i uczeniu maszynowym. Pozwala na zrozumienie i modelowanie związków między danymi, co jest niezbędne w wielu dziedzinach nauki i inżynierii. Umożliwia również przewidywanie wartości zmiennych zależnych na podstawie obserwacji zmiennych niezależnych, co ma szerokie zastosowanie między innymi w prognozowaniu ekonomicznym, analizie trendów, badaniach rynku czy medycynie.

Pomimo swojej użyteczności, regresja ma też swoje ograniczenia. Należy do nich założenie o liniowości związku między zmiennymi w regresji liniowej, co nie zawsze może być adekwatne do natury danych. W przypadku regresji logistycznej, choć model jest bardziej elastyczny, interpretacja współczynników i zrozumienie modelu może być bardziej skomplikowane. Ponadto, oba modele wymagają starannego rozważenia możliwości nadmiernego dopasowania, zwłaszcza w przypadku dużych zbiorów danych z wieloma zmiennymi [24, 39].

2.4. Problem grupowania

Grupowanie jest jedną z podstawowych technik uczenia nienadzorowanego w uczeniu maszynowym, której celem jest odkrywanie naturalnych podziałów w zestawie danych na podstawie podobieństwa między obiektami. W przeciwieństwie do uczenia nadzorowanego, grupowanie nie korzysta z etykietowanych danych, co oznacza, że algorytmy muszą samodzielnie identyfikować przynależność danych obiektów do klastrów czy grup [38].

Podstawą grupowania jest miara podobieństwa, która może być różna w zależności od rodzaju danych i celu analizy. Przykładem mogą być miary stosowane do obliczania odległości pomiędzy dwoma punktami. Aby tego dokonać można wykorzystać na przykład miarę odległości Euklidesowej (2.8) czy odległości Manhattan (2.9).

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{2.8}$$

$$d(p, q) = \sum_{i=1}^{n} |p_i - q_i| \tag{2.9}$$

gdzie p i q określają punkty, pomiędzy którymi odległość jest obliczana, w przestrzeni n-wymiarowej.

Innym podejściem do problemu grupowania jest wykorzystanie algorytmu K-means. Algorytm K-means to jeden z najprostszych i najczęściej stosowanych algorytmów grupowania. Polega na podziale zbioru danych na K klastrów, a następnie minimalizacji sumy kwadratów odległości między punktami a centroidami ich klastrów. Funkcję kosztu algorytmu K-means można zapisać jako:

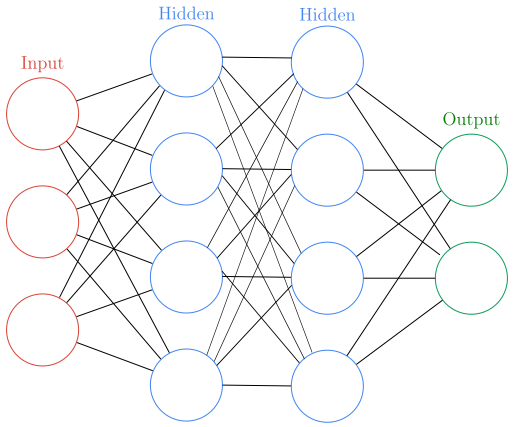
$$J = \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2.10}$$

gdzie S_i jest i-tym klastrem, x jest punktem należącym do klastra S_i, a μ_i jest centroidem klastra S_i [38].

2.5. Sieci neuronowe

Sieci neuronowe, stanowią kluczowy segment uczenia maszynowego, jednocześnie wywierając znaczący wpływ na postęp w dziedzinie sztucznej inteligencji. Jako systemy inspirowane strukturą i funkcjonowaniem ludzkiego mózgu, sieci neuronowe naśladują sposób, w jaki ludzkie neurony przetwarzają i interpretują informacje. Podstawową jednostką obliczeniową w sieci neuronowej jest sztuczny neuron, który, podobnie jak jego biologiczny odpowiednik, odbiera sygnały, przetwarza je i przekazuje dalej [1, 70].

Struktura sieci neuronowych składa się z warstw: warstwy wejściowej, która odbiera dane, jednej lub więcej warstw ukrytych, które przetwarzają dane, oraz warstwy wyjściowej, generującej wynik. W procesie uczenia sieci, wagi połączeń między neuronami są dostosowywane, co umożliwia modelowi adaptację i naukę z dostępnych danych. Rysunek 2.3 zawiera przykładową sieć neuronową z trzema neuronami na wejściu, dwoma na wyjściu i dwiema ukrytymi warstwami z 4 neuronami w każdej.



Rysunek 2.3: Przykładowa struktura sieci neuronowej

Rozwój sieci neuronowych został znacząco przyspieszony poprzez wzrost dostępnej mocy obliczeniowej i zwiększenie dostępności dużych zbiorów danych. Umożliwiło to skuteczne trenowanie modeli głębokiego uczenia, które są w stanie rozpoznawać wzorce i charakterystyki w danych, niedostępne dla tradycyjnych metod uczenia maszynowego. Głębokie sieci neuronowe, charakteryzujące się wieloma warstwami ukrytymi, znalazły zastosowanie w wielu problemach, takich jak przetwarzanie języka naturalnego czy rozpoznawanie obrazów, demonstrując swoją wszechstronność i skuteczność.

Kluczowym aspektem działania sieci neuronowych jest ich zdolność do generalizacji, czyli umiejętność modelu do poprawnego funkcjonowania na nowych, niewidzianych wcześniej danych. Aby osiągnąć wysoki poziom generalizacji, sieci są trenowane przy użyciu technik takich jak walidacja krzyżowa, która pomaga w ocenie, jak dobrze model będzie działał na innych, niedostępnych podczas trenowania danych.

Proces uczenia sieci neuronowych polega na dostosowywaniu wag połączeń między neuronami na podstawie danych wejściowych i oczekiwanych wyjść. To uczenie może odbywać się pod nadzorem, bez nadzoru lub w trybie wzmacniającym, w zależności od charakteru zadania i dostępnych danych. Kluczowym aspektem jest optymalizacja funkcji kosztu, która

mierzy różnicę między aktualnymi a oczekiwanymi wyjściami sieci. Sposoby optymalizacji oraz funkcje kosztu zostały szerzej opisane w kolejnych podrozdziałach 2.5.2, 2.5.3.

Wartość pojedynczego neuronu w sieci neuronowej jest zwykle wyliczana jako ważona suma jego wejść, do której dodaje się bias, a następnie wynik jest przekazywany przez funkcję aktywacji. Matematycznie, wartość neuronu y może być zapisany jako:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.11}$$

gdzie:

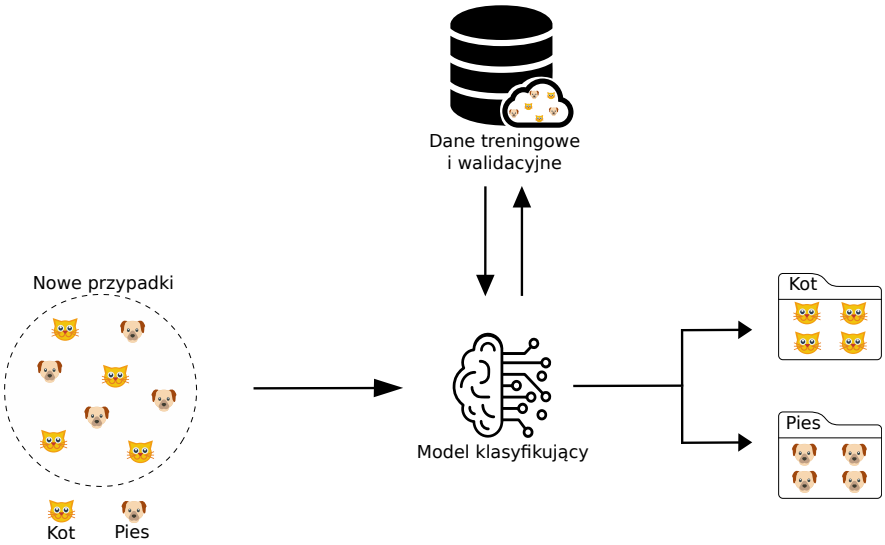
- f to funkcja aktywacji,
- n to liczba wejść do neuronu,
- w_i to waga i-tego wejścia,
- x_i to i-te wejście,
- b to bias neuronu.

Zastosowanie funkcji aktywacji f w neuronach jest fundamentalne dla działania sieci. Funkcje te decydują, czy dany neuron zostanie aktywowany, czyli czy i w jakim stopniu będzie przekazywał sygnał do następnej warstwy. Wybór odpowiedniej funkcji aktywacji ma zasadnicze znaczenie dla skuteczności modelu. Funkcje aktywacji wraz z przykładami zostały szerzej opisane w podrozdziale 2.5.1.

Sieci neuronowe są powszechnie stosowane w wizji komputerowej. Dane wizyjne stosowane są do wnioskowania oraz automatycznej analizy przechwyconych scen i uzyskiwania cennych informacji, podobnie jak robi to ludzkie oko. Wizja komputerowa jest szeroko stosowana w klasyfikacji obrazów lub wykrywaniu obiektów. W klasyfikacji obrazów algorytmy uczenia maszynowego uczą się przewidywać, co znajduje się na obrazie. Tymczasem wykrywanie obiektów automatycznie uzyskuje współrzędne interesujących obiektów na obrazie, na przykład w ruchu drogowym w celu liczenia samochodów na autostradach lub wykrywania pieszych w samochodach autonomicznych [9]. Przykładem klasyfikacji obrazu może być schemat na rysunku 2.4 gdzie to klasyfikator binarny ma za zadanie sklasyfikować obraz, czy jest na nim kot, czy pies.

Mimo wielu zalet, sieci neuronowe mają również swoje wyzwania. Jednym z nich jest ryzyko nadmiernego dopasowania (ang. overfitting), gdy model zbyt dokładnie dopasowuje się do danych treningowych kosztem zdolności do generalizacji. Problem nadmiernego dopasowania oraz dostępne techniki radzenia sobie z nim zostały przedstawione w podrozdziale 2.5.4.

W kontekście naukowym i praktycznym, rozumienie i rozwijanie sieci neuronowych wymaga interdyscyplinarnego podejścia, łączącego wiedzę z dziedziny informatyki, matematyki oraz statystyki. Tylko przez takie połączenie możliwe jest pełne wykorzystanie potencjału sieci neuronowych do rozwiązywania rzeczywistych problemów.



Rysunek 2.4: Przykład problemu klasyfikacji obrazu

2.5.1. Funkcje aktywacji

Funkcje aktywacji odgrywają kluczową rolę w sieciach neuronowych, umożliwiając im uczenie się i modelowanie złożonych, nieliniowych problemów. Są one istotnym elementem każdego neuronu w sieci, decydując o tym, czy i w jakim stopniu neuron przekazuje sygnał dalej do kolejnych warstw sieci. W tym podrozdziale przedstawiono zasadę działania funkcji aktywacji, ich roli w sieciach neuronowych oraz przedstawiono przykłady najpopularniejszych funkcji wraz z ich matematycznymi wzorami [92, 104, 136].

Funkcja aktywacji w neuronie przyjmuje ważoną sumę wejść neuronu plus bias jako argument (jak przedstawiono na wzorze (2.11)) i przekształca ten sygnał wejściowy w sygnał wyjściowy, który jest następnie przekazywany do kolejnej warstwy. Bez funkcji aktywacji, każdy neuron w sieci działałby jako prosty liniowy klasyfikator, co ograniczałoby zdolność sieci do modelowania tylko liniowych relacji między danymi wejściowymi a wyjściowymi. Dzięki wprowadzeniu nieliniowości poprzez funkcje aktywacji, sieci neuronowe mogą uczyć się i modelować znacznie bardziej złożone wzorce.

Rola funkcji aktywacji w sieci neuronowej jest wielowymiarowa. Po pierwsze, wprowadza nieliniowość do procesu przetwarzania danych, co jest niezbędne do efektywnego uczenia się skomplikowanych wzorców i zależności w danych. Po drugie, pomaga w regulacji przepływu informacji w sieci, aktywując neurony tylko wtedy, gdy jest to potrzebne, co zwiększa efektywność i stabilność uczenia. Po trzecie, różne funkcje aktywacji mogą być stosowane w różnych warstwach sieci, w zależności od specyficznych wymagań, co umożliwia większą elastyczność w projektowaniu architektury sieci.

W literaturze przedstawiono kilka podstawowych funkcji aktywacji [92]. Są to przede wszystkim:

- funkcja schodowa (ang. step activation function) jest jednym z najprostszych typów funkcji aktywacji. Jej działanie polega na zwróceniu stałej wartości (zazwyczaj 1) dla wszystkich wejść powyżej pewnego progu, i innej stałej wartości (zazwyczaj 0) dla

wartości poniżej tego progu. Matematycznie, funkcja ta może być zdefiniowana jako:

$$f(x) = \begin{cases} 1 & x > \text{próg} \\ 0 & x \leqslant \text{próg} \end{cases} \tag{2.12}$$

gdzie $f(x)$ to wartość funkcji aktywacji dla danego wejścia x, a *próg* to wartość graniczna, która decyduje o aktywacji neuronu. Jeśli suma wejściowa przekracza próg, neuron jest aktywowany (zwraca 1), w przeciwnym razie pozostaje nieaktywny (zwraca 0).

Funkcja schodkowa jest przykładem funkcji nieliniowej, która wprowadza jasny podział (decyzję) między dwoma stanami. Pomimo swojej prostoty, rzadko znajduje zastosowanie w praktyce poza bardzo prostymi modelami sieci neuronowych, głównie z powodu braku zdolności do modelowania złożonych zależności w danych.

- Funkcja liniowa, w kontekście sieci neuronowych, to funkcja aktywacji, która nie dokonuje żadnej transformacji wejść poza przeskalowaniem ich o stały współczynnik. Jest to najprostsza forma funkcji aktywacji, która zachowuje liniowość modelu. Funkcja liniowa jest zdefiniowana jako:

$$f(x) = ax + b \tag{2.13}$$

gdzie a jest współczynnikiem nachylenia, a b jest wyrazem wolnym.

Choć funkcja liniowa może wydawać się atrakcyjna ze względu na prostotę obliczeniową i interpretowalność, jej główna wada leży w niezdolności do modelowania złożonych zależności w danych. Sieć neuronowa, która zawiera tylko liniowe funkcje aktywacji (lub w ogóle ich nie stosuje), niezależnie od liczby warstw, będzie nadal ekwiwalentna sieci z pojedynczą warstwą z punktu widzenia możliwości aproksymacji funkcji. Dlatego też, w praktyce, funkcje liniowe są rzadko stosowane w ukrytych warstwach sieci neuronowych, choć mogą znajdować zastosowanie w warstwach wyjściowych dla zadań regresji.

Zarówno funkcja schodkowa jak i liniowa mają ograniczone zastosowanie w zaawansowanych strukturach sieci neuronowych, głównie z powodu ich prostoty i ograniczeń w modelowaniu złożonych zależności w danych. Dlatego w literaturze pojawiły się bardziej zaawansowane funkcje aktywacji, które przyczyniły się do postępu w dziedzinie sieci neuronowych:

- funkcja sigmoidalna (logistyczna) przekształca wartości wejściowe na zakres od 0 do 1, co sprawia, że jest przydatna w warstwach wyjściowych sieci realizujących zadania klasyfikacji binarnej i można ją opisać jako:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.14}$$

gdzie e to stała Eulera podniesiona do potęgi $-x$, gdzie x jest ważoną sumą wejść neuronu.

- Funkcja tangensu hiperbolicznego (tanh) podobnie jak sigmoid, wprowadza nieliniowość, ale jej zakres wyjściowy wynosi od -1 do 1, co często prowadzi do lepszej zbieżności modelu. Samą funkcję można zapisać następująco:

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.15}$$

- Funkcja ReLU (ang. rectified linear unit) jest obecnie jedną z najpopularniejszych funkcji aktywacji dzięki swojej prostocie obliczeniowej i skuteczności w wielu zastosowaniach. Promuje rzadszą aktywację neuronów, co może przyczynić się do efektywniejszego uczenia, zapisywana jest jako:

$$f(x) = \begin{cases} x & x > 0 \\ 0 & x \leqslant 0 \end{cases} \tag{2.16}$$

Jednym z najważniejszych aspektów stosowania funkcji aktywacji jest ich wpływ na proces uczenia sieci. Na przykład, funkcja ReLU i jej warianty, takie jak Leaky ReLU czy Parametric ReLU, pomagają w zarządzaniu problemem zanikających gradientów, który może utrudniać efektywne uczenie się głębokich sieci neuronowych.

W kontekście sieci wielowarstwowych, szczególnie ważne jest zastosowanie różnych funkcji aktywacji w różnych warstwach sieci. Dla przykładu, w warstwach ukrytych często stosuje się funkcje nieliniowe takie jak ReLU, aby umożliwić sieci modelowanie złożonych relacji. Z kolei na wyjściu sieci, w zależności od rodzaju zadania, mogą być stosowane takie funkcje jak softmax dla klasyfikacji wieloklasowej, której celem jest przypisanie prawdopodobieństw do poszczególnych klas wynikowych.

Kluczową cechą funkcji softmax jest jej zdolność do konwersji wektorów wartości (logitów) na rozkład prawdopodobieństwa, co jest szczególnie użyteczne w zadaniach klasyfikacyjnych, gdzie interesuje nas określenie stopnia pewności przynależności danej obserwacji do poszczególnych klas. Softmax gwarantuje, że wyjściowe prawdopodobieństwa dla wszystkich klas sumują się do jedności, co ułatwia interpretację wyników.

Wdrażając sieci neuronowe, ważne jest zrozumienie, jak wybór funkcji aktywacji wpływa na zdolność modelu do uczenia się i generalizacji. Eksperymentowanie z różnymi funkcjami aktywacji oraz ich parametrami może być kluczem do optymalizacji wydajności sieci dla konkretnego zadania. Ponadto, należy być świadomym potencjalnych problemów, takich jak na przykład zanikający gradient, które mogą wystąpić w trakcie uczenia, i wiedzieć, jak stosowanie określonych funkcji aktywacji może pomóc w ich rozwiązaniu.

Podsumowując, funkcje aktywacji są nieodłącznym elementem sieci neuronowych, odgrywającym kluczową rolę w umożliwianiu modelom uczenia się złożonych, nieliniowych wzorców. Wybór odpowiedniej funkcji aktywacji jest zatem fundamentalnym aspektem projektowania i optymalizacji sieci neuronowych, mającym bezpośredni wpływ na ich zdolność do rozwiązywania różnorodnych problemów.

2.5.2. Funkcje straty

Funkcje straty (ang. loss functions) są niezbędne do obliczania błędu sieci neuronowych. Funkcje straty odgrywają kluczową rolę w procesie uczenia, pozwalając na ocenę, jak bardzo przewidywania modelu różnią się od rzeczywistych wartości lub etykiet. Proces uczenia sieci neuronowej polega na minimalizacji tych strat poprzez iteracyjne dostosowywanie wag i biasów w modelu. W tym kontekście, funkcje straty są nie tylko miarą dokładności modelu, ale również kierunkowskazem, który prowadzi proces optymalizacji w kierunku poprawy modelu [36, 129, 132].

Funkcje straty można podzielić na różne kategorie w zależności od typu zadania, które sieć ma realizować - na przykład funkcje straty dla klasyfikacji i regresji. W kontekście klasyfikacji, jedną z najczęściej stosowanych funkcji straty jest kategoryczna entropia krzyżowa (ang. categorical cross-entropy loss), która jest szczególnie przydatna, gdy wyjście modelu jest interpretowane jako rozkład prawdopodobieństwa klas.

Funkcja kategorycznej entropii krzyżowej porównuje rozkład prawdopodobieństwa generowany przez model dla każdej klasy z rzeczywistym rozkładem, gdzie prawdziwa klasa ma prawdopodobieństwo 1, a pozostałe 0. Matematycznie, funkcja ta jest zdefiniowana jako:

$$L_i = -\sum_j y_{ij} \log(\hat{y}_{ij}) \tag{2.17}$$

gdzie:

- L_i to strata dla i-tej próbki,
- j to indeks klasy,
- y_{ij} to prawdziwa etykieta dla i-tej próbki i j-tej klasy,
- $\hat{y}_{ij}$ to przewidywane prawdopodobieństwo, że i-ta próbka należy do j-tej klasy.

Funkcja ta skutecznie karze model za nadmierne przypisywanie wysokiego prawdopodobieństwa niewłaściwym klasom, wspierając tym samym zwiększoną pewność w przypadku prawidłowych klasyfikacji. Dlatego entropia krzyżowa stanowi cenny instrument w zadaniach klasyfikacyjnych, gdzie priorytetem jest nie tylko dokonanie właściwego przyporządkowania klasy, ale także zapewnienie wysokiego stopnia pewności tej klasyfikacji.

W przypadku zadań regresji często stosowaną funkcją straty jest błąd średniokwadratowy (ang. mean squared error, MSE), który mierzy średnią kwadratów różnic między przewidywanymi wartościami, a rzeczywistymi wartościami. MSE można zapisać jako:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2.18}$$

gdzie:

- n to liczba próbek,
- $\hat{y}_i$ to przewidywana wartość dla i-tej próbki,
- y_i to rzeczywista wartość dla i-tej próbki.

Funkcja straty MSE jest szczególnie użyteczna w sytuacjach, gdy ważne jest karanie dużych błędów bardziej niż małych, co wynika z kwadratowej natury funkcji.

Funkcje strat w sieciach neuronowych pełnią kluczową rolę w procesie uczenia, dostarczając miarę błędu pomiędzy przewidywaniami modelu, a rzeczywistymi etykietami danych. Są one fundamentem mechanizmu uczenia się, ponieważ pozwalają algorytmowi optymalizacyjnemu zrozumieć, w jaki sposób powinien dostosować wagi w sieci, aby zmniejszyć błąd przewidywań.

Funkcja straty mierzy, jak bardzo model myli się w swoich przewidywaniach. Idealnie aby strata była jak najmniejsza, a idealnie równa zero, co oznaczałoby, że przewidywania

modelu są w pełni zgodne z rzeczywistymi etykietami danych. Jednak w praktyce, zwłaszcza na początkowych etapach uczenia, model często generuje przewidywania dalekie od idealnych, co skutkuje większymi wartościami strat.

Kategoryczna entropia krzyżowa jest szczególnie efektywna w kontekście modeli stosujących funkcję aktywacji softmax na wyjściowej warstwie, gdzie przewidywania są prezentowane jako rozkłady prawdopodobieństwa przynależności do poszczególnych klas. Poprzez porównanie rozkładu prawdopodobieństw przewidywanych przez model z rzeczywistym rozkładem etykiet, entropia krzyżowa efektywnie ocenia, jak dobrze model radzi sobie z przewidywaniem prawidłowych klas. Z kolei średni błąd kwadratowy (MSE) mierzy średnią kwadratów różnic między wartościami przewidywanymi, a rzeczywistymi będąc standardowym wyborem w problemach regresji, gdzie celem jest przewidywanie ciągłych wartości.

Podsumowując, funkcje straty są nieodłącznym elementem procesu uczenia sieci neuronowych, umożliwiając nie tylko ocenę błędu modelu, ale również kierując procesem jego optymalizacji. Dokonując świadomego wyboru funkcji straty, zależnego od rodzaju problemu i specyfiki danych, możliwe jest zwiększenie efektywności uczenia.

2.5.3. Optymalizatory

Optymalizatory są niezbędnym elementem procesu uczenia sieci neuronowych. Optymalizatory odpowiadają za aktualizację wag i biasów w sieci w oparciu o obliczone gradienty, aby minimalizować funkcję strat. Proces ten jest kluczowy dla efektywnego uczenia modeli i poprawy ich dokładności [32, 51, 126].

Optymalizator bazujący na stochastycznym zejściu gradientu (ang. stochastic gradient descent, SGD) jest jednym z najprostszych, ale zarazem fundamentalnych optymalizatorów. Podstawowa idea polega na aktualizacji parametrów modelu przez odjęcie gradientu funkcji straty pomnożonego przez stałą zwaną szybkością uczenia (ang. learning rate). Chociaż jest to podejście podstawowe, różne warianty optymalizatora SGD są szeroko stosowane ze względu na ich skuteczność i prostotę implementacji.

Adaptacyjny algorytm gradientowy (ang. adaptive gradient algorithm, AdaGrad) to kolejny optymalizator, który wprowadza indywidualną szybkość uczenia poprzez dostosowanie jej na podstawie historii gradientów. Pozwala to na bardziej zrównoważone aktualizacje, szczególnie w przypadku rzadko występujących cech. Choć AdaGrad skutecznie radzi sobie z problemami o dużej skali, jego ciągłe akumulowanie kwadratów gradientów może prowadzić do zbyt szybkiego malejącego tempa uczenia się.

Algorytm propagacji pierwiastka średniej kwadratowej (ang. root mean square propagation, RMSProp) jest usprawnieniem optymalizatora AdaGrad, które rozwiązuje problem zbyt szybkiego spadku tempa uczenia poprzez wprowadzenie zanikającej średniej ruchomej kwadratów gradientów. Dzięki temu RMSProp zachowuje adaptacyjność AdaGrad, jednocześnie unikając pułapki zbyt małej szybkości uczenia w późniejszych etapach treningu. RMSProp jest często używany w praktyce i rekomendowany w wielu zastosowaniach uczenia głębokiego.

Optymalizator adaptacyjnej estymacji momentu (ang. adaptive moment estimation, Adam) jest zaawansowanym optymalizatorem, który łączy w sobie idee dwóch kluczowych koncepcji: adaptacyjnych szybkości uczenia się oraz momentu, co sprawia, że jest wyjątkowo skuteczny w różnorodnych zastosowaniach uczenia głębokiego. „Moment” w kontekście optymalizatorów odnosi się do mechanizmu uwzględniającego zarówno bieżące gradienty, jak i gradienty

z poprzednich kroków, w celu wygładzenia ścieżki aktualizacji parametrów modelu. Algorytm Adam stosuje tzw. moment pierwszego rzędu (średnią ruchomą gradientów) oraz moment drugiego rzędu (średnią ruchomą kwadratów gradientów), aby adaptacyjnie dostosować szybkość uczenia się dla każdego parametru modelu. Dzięki temu optymalizator Adam automatycznie dostosowuje wielkość kroków aktualizacji parametrów, co pomaga w szybszej i bardziej stabilnej generalizacji sieci neuronowej. Moment pierwszego rzędu pomaga w utrzymaniu kierunku aktualizacji, zmniejszając ryzyko utknięcia w lokalnych minimach, podczas gdy moment drugiego rzędu reguluje szybkość uczenia się, zapewniając indywidualne tempo dla każdego parametru. W praktyce oznacza to, że algorytm Adam jest bardzo efektywny w radzeniu sobie z problemami, które mają nieregularne skale w danych.

Wybór odpowiedniego optymalizatora i dostosowanie jego hiperparametrów, takich jak szybkość uczenia czy wartości momentu, ma zasadnicze znaczenie dla osiągnięcia wysokiej wydajności modelu sieci neuronowej. Eksperymentowanie z różnymi optymalizatorami i ich ustawieniami jest kluczowym elementem procesu trenowania modeli, pozwalającym na znalezienie najlepszego rozwiązania dla danego zadania.

2.5.4. Problem nadmiernego dopasowania modelu w sieciach konwolucyjnych

Badania nad konwolucyjnymi sieciami neuronowymi koncentrują się w dużej mierze na ich zdolności do uogólniania. Problem szczególnie pojawia się, gdy model CNN staje się nadmiernie skomplikowany, na przykład posiada zbyt dużą liczbę parametrów w stosunku do dostępnej liczby próbek szkoleniowych. Taki model może ulec zjawisku znanemu jako nadmierne dopasowanie, co prowadzi do osłabienia zdolności do generalizacji. W rezultacie, model taki może zaczynać reprezentować losowy błąd lub szum w danych, zamiast odzwierciedlać ich podstawowy rozkład. W skrajnych przypadkach model CNN może wykazywać wysoką wydajność na danych treningowych, lecz zawodzić przy przewidywaniach na nowych, nieznanych wcześniej danych [127, 130].

Nadmierne dopasowanie stanowi poważne wyzwanie w dziedzinie uczenia maszynowego oraz konwolucyjnych sieci neuronowych (CNN). Strategie minimalizacji ryzyka nadmiernego dopasowania można podzielić na dwie główne kategorie: regularyzację (ang. regularization) oraz rozszerzanie danych (ang. data augmentation) [130].

Regularyzacja odgrywa fundamentalną rolę w zapobieganiu nadmiernemu dopasowaniu w procesie uczenia modeli konwolucyjnych sieci neuronowych (CNN). W literaturze naukowej zaproponowano wiele różnych metod regularyzacji, które mają na celu zwiększenie efektywności i ogólnej zdolności generalizacyjnej tych sieci [3, 43, 58, 113, 123, 126]. Spośród nich można wyszczególnić następujące metody:

- normalizacja danych (ang. data normalization) - metoda polegająca na normalizacji danych wejściowych co poprawia stabilność i wydajność sieci [37]. W przypadku przetwarzania obrazów dokonuje się normalizacji wartości pikseli z przedziału $[0, 255]$ do przedziału $[0, 1]$.

- Dropout - jest to metoda polegająca na losowym wyłączaniu (ustawianiu na zero) wyjścia poszczególnych ukrytych neuronów z określonym prawdopodobieństwem pod-

czas treningu. Pozwala to na zmniejszenie zależności między neuronami i zapobiega nadmiernemu dopasowaniu [58, 97].

- DropConnect - jest to uogólnienie metody Dropout. W tym podejściu losowo wybierane są wagi, które zostają ustawione na zero podczas treningu, co również zmniejsza ryzyko nadmiernego dopasowania [113].

- Adaptive Dropout - w tej metodzie prawdopodobieństwo wyłączenia każdego ukrytego neuronu jest szacowane za pomocą binarnej sieci przekonań, co umożliwia bardziej dynamiczne dostosowanie procesu uczenia [3].

- Stochastic Pooling - metoda ta polega na losowym wyborze aktywacji z rozkładu wielomianowego podczas treningu, jest to metoda bez parametryczna i może być stosowana w połączeniu z innymi technikami regularyzacji [126].

- DisturbLabel - ta technika wprowadza szum do warstwy kosztu poprzez losową zmianę etykiet małej części próbek na nieprawidłowe wartości podczas każdej iteracji treningu [123].

- PatchShuffle - metoda ta polega na losowym przemieszczaniu pikseli w każdym lokalnym elemencie obrazu, co pozwala zachować globalne struktury, ale wprowadza lokalne wariacje, co może być korzystne dla treningu sieci CNN [43].

Każda z powyższych metod wprowadza innowacyjne podejścia do regularyzacji, które mają na celu poprawę zdolności modelu CNN do efektywnego i generalizującego uczenia się.

Rozszerzanie danych jest uznawaną metodą regularyzacji, powszechnie stosowaną w procesie uczenia konwolucyjnych sieci neuronowych (CNN) [34, 58, 94]. Technika ta ma na celu sztuczne zwiększenie objętości zbioru danych treningowych poprzez zastosowanie różnych realistycznych transformacji obrazów.

Należą do nich:

- losowe przycinanie (ang. random cropping): technika polega na na wyodrębnieniu losowej części obrazu wejściowego, co pomaga w modelowaniu różnorodności w położeniu i skali obiektów [58].

- Losowe obracanie (ang. random rotate): operacja polega na losowym obracaniu obrazów podczas treningu, np. o 13° w lewo [58].

- Losowe odwracanie (ang. random flipping): Metoda polegająca na odwracaniu obrazów w poziomie lub pionie, co wprowadza dodatkową różnorodność w orientacji obiektów [94].

Techniki rozszerzania danych są kluczowe w zwiększaniu zdolności modeli CNN do generalizacji i zapobieganiu ich nadmiernemu dopasowaniu. Ma to również pozytywne przełożenie na lepszą wydajność przy pracy z nowymi, nieznajomymi danymi. Przykładowe przekształcenia w procesie rozszerzania danych przedstawia rysunek 2.5.

Rysunek 2.5: Przykładowe przekształcenia w procesie rozszerzania danych

2.6. Ocena jakości klasyfikacji

W niniejszym podrozdziale zostały szczegółowo przedstawione i opisane metryki stosowane do oceny wydajności modeli uczenia maszynowego, ze szczególnym uwzględnieniem uczenia nadzorowanego. Ocena jakości klasyfikacji stanowi nierozłączny element procesu szkolenia każdego modelu uczenia maszynowego, pełniąc kluczową rolę w określeniu skuteczności modelu w przewidywaniu wyników na podstawie dostępnych danych. W kontekście uczenia nadzorowanego, gdzie model jest trenowany na podstawie danych zawierających zarówno wejścia, jak i oczekiwane wyjścia (etykiety), ocena jakości klasyfikacji umożliwia nie tylko weryfikację dokładności modelu, ale także identyfikację obszarów wymagających dalszej optymalizacji.

Ocena skuteczności modeli klasyfikacyjnych obejmuje szereg miar, które pozwalają na kompleksową analizę ich wydajności. Każda z tych miar, będących przedmiotem dokładnego opisu w kolejnych podrozdziałach, odnosi się do różnych aspektów jakości klasyfikacji, takich jak precyzja, czułość, dokładność czy wartość miary F1. W zależności od specyfiki wymagań i charakteru danych, różne miary mogą być stosowane do wyróżniania odmiennych cech modelu, co podkreśla konieczność zrównoważonego podejścia do oceny jakości klasyfikacji.

Ważnym aspektem oceny modeli klasyfikacyjnych jest macierz pomyłek, która dostarcza podstawowe informacje o liczbie poprawnych i błędnych klasyfikacjach dokonanych przez model. Na podstawie tej macierzy wyliczane są wspomniane wcześniej miary, takie jak precyzja (ang. precision), stosunek poprawnie pozytywnych predykcji do wszystkich pozytywnych predykcji dokonanych przez model czy czułość (ang. recall), stosunek poprawnie pozytywnych predykcji

do wszystkich faktycznych pozytywnych przypadków w danych. Te i inne miary, szczegółowo opisane w dalszej części rozdziału, pozwalają na wielowymiarową analizę wydajności modelu, uwzględniając zarówno jego mocne, jak i słabe strony.

Ponadto, w rozważaniach na temat oceny jakości klasyfikacji nie można pominąć kwestii nierównomiernie rozłożonych klas (tzw. niezbalansowanych danych), które mogą istotnie wpłynąć na interpretację wyników oceny. W takich sytuacjach tradycyjne miary, takie jak dokładność (ang. accuracy), mogą być mylące, co wymusza stosowanie bardziej zaawansowanych technik analizy wydajności, takich jak na przykład zbalansowana dokładność (ang. balanced accuracy)

W niniejszym rozdziale szczegółowo omówiono kluczowe miary oceny jakości klasyfikacji, prezentując zarówno ich zalety, jak i ograniczenia, a także podając wzory umożliwiające ich wyliczenie. Analiza ta stanowi fundament dla zrozumienia skuteczności modeli klasyfikacyjnych w uczeniu nadzorowanym, podkreślając ich znaczenie w kontekście rozwoju precyzyjnych i efektywnych systemów uczenia maszynowego [29, 69, 95].

2.6.1. Dokładność

Dokładność (ang. accuracy) jest jedną z najbardziej intuicyjnych i powszechnie stosowanych miar oceny modeli klasyfikacyjnych w uczeniu maszynowym, szczególnie w kontekście uczenia nadzorowanego. Definiuje się ją jako stosunek liczby poprawnych predykcji (zarówno pozytywnych, jak i negatywnych) do całkowitej liczby przypadków w zestawie danych. Wzór na dokładność wyraża się następująco:

$$accuracy = \frac{\sum_{i=1}^{k} \frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}}{k} \tag{2.19}$$

gdzie:

- k określa liczbę klas,
- i jest klasą generyczną, w zależności dla której klasy metryka jest wyliczana,
- TP_i są to prawdziwie pozytywne przypadki (ang. true positive), które określają liczbę prawidłowo sklasyfikowanych przypadków klasy i,
- TN_i są to prawdziwie negatywne przypadki (ang. true negative), które określają liczbę prawidłowo sklasyfikowanych przypadków jako nie klasa i,
- FP_i są to fałszywie pozytywne przypadki (ang. false positive), które określają liczbę niepoprawnie sklasyfikowanych przypadków klasy i,
- FN_i są to fałszywie negatywne przypadki (ang. false negative), które określają liczbę niepoprawnie sklasyfikowanych przypadków jako nie klasa i.

Zaletą dokładności jako miary jest jej prostota i bezpośrednia interpretowalność. Umożliwia szybką ocenę ogólnej wydajności modelu, dostarczając podstawowych informacji na temat jego skuteczności w klasyfikacji. Jest to szczególnie przydatne w przypadkach, gdy mamy do czynienia z równomiernie rozłożonymi klasami w danych, na których klasyfikator był uczony.

Dokładność, choć jest miarą intuicyjną i szeroko stosowaną, może nie zawsze najlepiej odzwierciedlać rzeczywistą wydajność modelu klasyfikacyjnego, zwłaszcza w specyficznych scenariuszach zastosowań. Główną wadą dokładności jest jej potencjalne wprowadzanie w błąd oceny modeli w przypadku niezbalansowanych zbiorów danych. Przykładowo, w sytuacji, gdzie 90% próbek należy do jednej klasy, model zawsze przewidujący tę dominującą klasę osiągnie dokładność na poziomie 90%, mimo że nie wykazuje żadnej zdolności do rozróżniania klas w bardziej zrównoważony sposób. W takich przypadkach, inne miary, takie jak precyzja, czułość czy miara F1, mogą dostarczyć bardziej kompleksowego obrazu wydajności modelu, uwzględniając jego zdolność do prawidłowego klasyfikowania próbek z poszczególnych klas.

W kontekście naukowym, przy omawianiu dokładności oraz innych miar oceny, kluczowe jest uwzględnienie charakterystyki zbioru danych i specyfiki problemu, który ma zostać rozwiązany. Na przykład, w aplikacjach medycznych, gdzie koszt fałszywie negatywnej klasyfikacji (FN) – pominięcia istotnej diagnozy – może być bardzo wysoki, czułość staje się ważniejsza niż dokładność. W takich przypadkach, model z niższą ogólną dokładnością, ale wyższą czułością, może być bardziej pożądany.

Podsumowując, choć dokładność jest ważną i często stosowaną miarą oceny modeli klasyfikacyjnych, skuteczność tej miary jako wskaźnika wydajności może być ograniczona w specyficznych kontekstach. Właściwe zrozumienie i stosowanie dokładności, w połączeniu z innymi miarami, pozwala na bardziej kompleksową i precyzyjną ocenę modeli klasyfikacyjnych. To podkreśla znaczenie holistycznego podejścia do oceny wydajności w uczeniu maszynowym, gdzie różne miary są analizowane łącznie, aby dostarczyć pełny obraz możliwości modelu.

2.6.2. Precyzja

Precyzja (ang. precision) jest kluczową miarą w ocenie modeli klasyfikacyjnych, zwłaszcza w kontekstach, gdzie istotne jest minimalizowanie liczby fałszywie pozytywnych wyników. Precyzja określa proporcję prawdziwie pozytywnych predykcji względem całkowitej liczby predykcji uznanych przez model za pozytywne. Innymi słowy, mierzy ona dokładność modelu wśród przypadków, które zostały zaklasyfikowane jako pozytywne. Wzór na precyzję można zapisać jako:

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{2.20}$$

Jedną z głównych zalet precyzji jest zdolność do oceny efektywności modelu w ograniczaniu liczby błędnych pozytywnych alarmów. Jest to szczególnie ważne w zastosowaniach, gdzie koszty fałszywie pozytywnych wyników są wysokie, na przykład w medycynie, gdzie fałszywie pozytywny wynik może skutkować niepotrzebnymi lub inwazyjnymi procedurami diagnostycznymi, czy w systemach rekomendacyjnych, gdzie nadmierne polecanie nieodpowiednich produktów może prowadzić do frustracji użytkowników.

Precyzja jest jednak miarą, która nie bierze pod uwagę wszystkich aspektów wydajności modelu. Skupiając się wyłącznie na proporcji prawidłowo zidentyfikowanych pozytywnych przypadków wśród wszystkich przypadków uznanych za pozytywne, ignoruje ona te prawdziwie pozytywne przypadki, które zostały przeoczone (FN, False Negatives). Dlatego precyzja może być wysoka, nawet jeśli model pomija znaczącą liczbę pozytywnych przypadków, co nie jest

idealne w sytuacjach, gdzie ważne jest wykrycie jak największej liczby takich przypadków, np. w diagnostyce chorób.

Z tego powodu precyzja często jest używana razem z miarą czułości (ang. recall), aby zapewnić bardziej zrównoważoną ocenę modelu. Łączną ocenę precyzji oraz czułości umożliwia miara F1, która jest harmoniczną średnią obu tych miar i dostarcza pojedynczą metrykę uwzględniającą zarówno potrzebę unikania fałszywie pozytywnych wyników, jak i konieczność identyfikacji wszystkich pozytywnych przypadków.

Precyzja jest niezwykle ważną miarą w kontekstach, gdzie kluczowe jest zminimalizowanie fałszywie pozytywnych wyników, jednak jej użyteczność wzrasta, gdy jest stosowana w połączeniu z innymi miarami, takimi jak czułość, co pozwala na uzyskanie pełniejszego obrazu wydajności modelu klasyfikacyjnego. Takie wieloaspektowe podejście do oceny modeli klasyfikacyjnych jest niezbędne do dokładnego zrozumienia ich możliwości i ograniczeń.

2.6.3. Czułość

Czułość (ang. recall), znana również jako wskaźnik prawdziwie pozytywnych przypadków (ang. True Positive Rate, TPR), jest miarą oceny wydajności modeli klasyfikacyjnych, która koncentruje się na zdolności modelu do prawidłowego identyfikowania wszystkich przypadków klasy pozytywnej. W kontekście uczenia maszynowego, czułość określa proporcje wszystkich rzeczywistych pozytywnych przypadków, które zostały poprawnie zidentyfikowane przez model jako pozytywne. Wzór na czułość wyraża się następująco:

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{2.21}$$

Zaletą miary czułości jest jej skupienie na kluczowym aspekcie klasyfikacji w wielu zastosowaniach: zdolności do identyfikacji wszystkich pozytywnych przypadków. Jest to szczególnie ważne w problemach, gdzie niezidentyfikowanie pozytywnego przypadku może mieć poważne konsekwencje, na przykład w diagnostyce medycznej, gdzie pominięcie pozytywnej diagnozy (np. nowotworu) jest bardziej niepożądane niż fałszywie pozytywna diagnoza. Czułość pozwala zatem na ocenę, jak dobrze model radzi sobie z wyłapywaniem wszystkich pozytywnych wyników, co jest kluczowe dla zapewnienia wysokiej jakości klasyfikacji w takich krytycznych zastosowaniach.

Jednakże, czułość nie jest pozbawiona wad. Jedną z głównych jest fakt, że skupia się wyłącznie na pozytywnych przypadkach, ignorując liczbę fałszywie pozytywnych (FP) wyników. Oznacza to, że model może mieć wysoką czułość, identyfikując większość lub wszystkie pozytywne przypadki, ale równocześnie generować dużą liczbę błędnych alarmów (FP), co w niektórych zastosowaniach może być nieakceptowalne. Na przykład, w systemach zabezpieczeń, duży odsetek fałszywych alarmów może prowadzić do ignorowania prawdziwych alertów przez operatorów.

Ponadto, wysoka czułość nie zawsze oznacza wysoką ogólną skuteczność modelu. Model może być bardzo dobry w identyfikacji pozytywnych przypadków, ale jeśli robi to kosztem znaczącej liczby fałszywych alarmów, jego użyteczność może być ograniczona. Dlatego czułość często analizuje się w połączeniu z innymi miarami, takimi jak na przykład precyzja (proporcja prawdziwie pozytywnych wyników wśród wszystkich wyników zidentyfikowanych jako pozytywne przez model) aby uzyskać pełniejszy obraz wydajności modelu.

Czułość jest niezwykle ważną miarą w kontekście oceny modeli klasyfikacyjnych, zwłaszcza tam, gdzie niezidentyfikowanie pozytywnego przypadku niesie za sobą poważne konsekwencje. Jednakże, jak każda miara, ma swoje ograniczenia i najlepiej sprawdza się, gdy jest stosowana w połączeniu z innymi metrykami, co pozwala na zrównoważoną ocenę wydajności modelu w różnych aspektach klasyfikacji.

2.6.4. Zbalansowana dokładność

Zbalansowana dokładność (ang. balanced accuracy) stanowi istotne rozszerzenie tradycyjnej miary dokładności, szczególnie przydatne w kontekście niezrównoważonych zbiorów danych. Tradycyjna dokładność, choć intuicyjna i szeroko stosowana, może wprowadzać w błąd w ocenie modeli, gdy rozkład klas jest w znacznej dysproporcji. Zbalansowana dokładność została wprowadzona, aby złagodzić ten problem, poprzez równomierne traktowanie wyników z różnych klas. Wzór na zbalansowaną dokładność jest średnią z czułości i specyficzności:

$$balanced\ accuracy = \frac{\sum_{i=1}^{k} recall_i + TNR_i}{k} \tag{2.22}$$

gdzie:

- TNR to wskaźnik prawdziwie negatywnych przypadków (ang. True Negative Rate) definiowany jako $\frac{TN}{TN+FP}$, znany również jako specyficzność lub selektywność.

Czułość jest miarą zdolności modelu do prawidłowego identyfikowania pozytywnych przypadków, a specyficzność mierzy zdolność do prawidłowego identyfikowania negatywnych przypadków.

Zaletą zbalansowanej dokładności jest jej odporność na problem niezrównoważonych zbiorów danych. Poprzez średnią z czułości i specyficzności, zbalansowana dokładność zapewnia bardziej odporną ocenę wydajności modelu, ponieważ każda klasa, niezależnie od jej liczności, ma równy wpływ na wynik. Daje to możliwość lepszego porównywania modeli, szczególnie w przypadkach, gdy zainteresowanie budzi zdolność modelu do równie skutecznego identyfikowania przypadków z różnych klas.

Jednakże zbalansowana dokładność również ma swoje ograniczenia. Choć lepiej radzi sobie w przypadku niezrównoważonych zbiorów danych niż tradycyjna dokładność, nie dostarcza informacji na temat konkretnej natury błędów klasyfikacji, takich jak fałszywie pozytywne i fałszywie negatywne wyniki. Może to być istotne w zastosowaniach, gdzie różne typy błędów niosą za sobą odmienne konsekwencje.

Zbalansowana dokładność oferuje ważną alternatywę dla tradycyjnej miary dokładności, szczególnie w kontekście analizy modeli działających na niezrównoważonych zbiorach danych. Dostarcza bardziej zrównoważonej i odpornej oceny, lepiej odzwierciedlając faktyczną wydajność modelu w różnych warunkach. Należy jednak pamiętać o jej ograniczeniach i rozważać jej stosowanie w kontekście szerszego zestawu metryk, aby uzyskać kompleksowy obraz wydajności modelu klasyfikacyjnego.

2.6.5. Miara F1

Miarę F1 można uznać za jedną z najbardziej istotnych metryk w kontekście oceny modeli klasyfikacyjnych, zwłaszcza gdy poszukuje się zrównoważonej oceny między precyzją, a czu-

łością. Miara F1 jest harmoniczną średnią precyzji i czułości, oferując pojedynczą metrykę, która bierze pod uwagę zarówno zdolność modelu do precyzyjnego identyfikowania pozytywnych przypadków, jak i jego skuteczność w wykrywaniu jak największej liczby prawdziwych przypadków pozytywnych. Wzór na miarę F1 wyraża się następująco:

$$F1_i = \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \tag{2.23}$$

Zaletą miary F1 jest jej zdolność do dostarczania zbilansowanego oglądu wydajności modelu, szczególnie w sytuacjach, gdzie rozkład klas jest niezrównoważony. Miara F1 staje się krytycznie ważna, gdy nie można przedkładać precyzji, ani czułości kosztem drugiego wskaźnika, a poszukuje się metodyki oceny, która scala oba te wskaźniki.

Miara F1, będąca harmonijną średnią precyzji i czułości, jest ceniona za zdolność do zrównoważonego oceniania modeli klasyfikacyjnych, lecz posiada także ograniczenia. Głównym z nich jest stałe traktowanie precyzji i czułości jako równie ważnych, co może nie oddawać rzeczywistych potrzeb specyficznych zastosowań. Na przykład, w zastosowaniach, gdzie konsekwencje fałszywie pozytywnych decyzji są bardziej kosztowne niż pominięcia prawdziwie pozytywnych przypadków, standardowa miara F1 może nie dostarczać pełni informacji o wydajności modelu. Co więcej, skupienie się wyłącznie na pozytywnych wynikach klasyfikacji pomija informacje o zdolności modelu do identyfikacji negatywnych przypadków, co w niektórych sytuacjach może być równie istotne.

W przypadkach, w których istnieje potrzeba wyróżnienia precyzji lub czułości, można zastosować ważoną wersję miary F1, znaną również jako F-beta (F_β), której wzór przedstawia się następująco:

$$F1_{i\beta} = (1 + \beta^2) \cdot \frac{precision_i \cdot recall_i}{(\beta^2 \cdot precision_i) + recall_i} \tag{2.24}$$

Parametr β pozwala na dostosowanie wagi przydzielanej czułości względem precyzji. Gdy β jest większe niż 1, większe znaczenie przykładane jest do czułości, co jest korzystne w sytuacjach, gdzie niezidentyfikowanie pozytywnego przypadku niesie za sobą poważniejsze konsekwencje. Z kolei wartość β mniejsza niż 1 kładzie nacisk na precyzję, co może być pożądane w zastosowaniach, gdzie ważniejsze jest zminimalizowanie liczby fałszywie pozytywnych wyników.

Dzięki możliwości dostosowania wagi między precyzją, a czułością miara F_β oferuje bardziej elastyczną możliwość oceny modeli klasyfikacyjnych, lepiej odpowiadając na specyficzne wymagania różnych zastosowań. Jednakże, podobnie jak w przypadku standardowej miary F1, ważne jest, aby pamiętać o ograniczeniach tej metryki.

Podsumowując, miara $F1$ oraz F_β jest niezastąpiona w wielu scenariuszach oceny modeli klasyfikacyjnych, zapewniając zrównoważony pomiar wydajności w sytuacjach wymagających uwzględnienia zarówno precyzji, jak i czułości. Niemniej jednak, zawsze należy dokonać wyboru metryki oceny w kontekście specyficznych wymagań, co może oznaczać potrzebę uzupełnienia analizy skuteczności modelu o dodatkowe miary.

3. Problematyka

Niniejszy rozdział stanowi wstęp do problematyki analizy zachowań zawodników w boksie olimpijskim. Przetwarzanie obrazów w kontekście sportowym, już od dłuższego czasu, znajduje szerokie zastosowanie w różnorodnych dyscyplinach, począwszy od piłki nożnej, poprzez koszykówkę, aż po tenis. Możliwości, jakie oferują obecne technologie wizyjne, pozwalają na automatyczne wydobywanie informacji z materiałów wideo, co ma kluczowe znaczenie nie tylko dla trenerów i analityków, ale również dla samych sportowców. Jednocześnie, dostępność i rozwój otwartych baz danych znacznie przyczyniają się do postępu w dziedzinie przetwarzania obrazu i uczenia maszynowego, umożliwiając rozwijanie nowych metod analizy oraz weryfikację istniejących rozwiązań.

W kontekście boksu olimpijskiego, analiza zachowań zawodników przybiera na znaczeniu, zwłaszcza w świetle możliwości, jakie oferuje zastosowanie technologii wizyjnych i uczenia maszynowego. Rozpoznawanie ciosów, ocena strategii oraz analiza ruchów zawodników na podstawie danych wizyjnych otwierają nowe perspektywy dla treningu, budowania strategii walki oraz oceny kondycji i technik zawodników. Osiągnięcie wysokiej skuteczności w tych obszarach wymaga jednak nie tylko zaawansowanych narzędzi, ale również dogłębnego zrozumienia specyfiki boksu olimpijskiego, w tym regulacji dotyczących oceny i ubioru.

Niniejszy rozdział ma na celu nie tylko przedstawienie aktualnych podejść i wyzwań związanych z analizą zachowań bokserów, ale również podkreślenie znaczenia, jakie mają dostępne bazy danych w kontekście badań nad przetwarzaniem obrazu w sporcie. Poprzez przegląd obecnych metod analizy oraz omówienie potencjału, jaki niosą za sobą dane wizyjne, rozdział ten wprowadza w problematykę badań skupionych na zastosowaniach technologii wizyjnych w boksie olimpijskim, zwracając uwagę na kluczowe aspekty i wyzwania, które należy uwzględnić przy opracowywaniu efektywnych metod analizy zachowań zawodników.

3.1. Przetwarzanie obrazu w sporcie

Niniejszy podrozdział poświęcony jest analizie zastosowania technologii wizyjnych w sporcie, zawierając przegląd badań i metod w różnych dyscyplinach sportowych. Rozwój technik rejestrowania obrazu i wizji komputerowej otwiera nowe możliwości dla oceny wydajności zawodników i zrozumienia dynamiki gry. Przedstawiane rozwiązania, dostosowane do specyficznych potrzeb różnych sportów, pozwalają na automatyczne wydobywanie kluczowych informacji z materiałów wideo, dostarczając cennych danych dla trenerów, analityków i widzów.

Stosowanie technologii wizyjnych w sportach takich jak piłka nożna, piłka ręczna, tenis czy koszykówka demonstruje ich wszechstronność i zdolność do dostarczania szczegółowych analiz. Od automatycznego podsumowywania wydarzeń z gry, przez dostarczanie dodatkowych informacji na temat akcji na boisku, aż po złożoną analizę wysokiego poziomu, techniki te umożliwiają głębsze zrozumienie aspektów technicznych i taktycznych gry. Integracja danych wideo z zaawansowanymi algorytmami przetwarzania obrazu pozwala na identyfikację

i klasyfikację zdarzeń, monitorowanie zmian położenia zawodników, a także ocenę ich postawy ciała i aktywności.

Przegląd ten podkreśla znaczący postęp w dziedzinie analizy sportowej dzięki zastosowaniu technologii wizyjnych, wskazując na ich potencjał do wzbogacenia treningu, strategii gry i rozwoju sportu. Poprzez ciągłe doskonalenie metod i rozwój nowych narzędzi, możliwe jest uzyskanie jeszcze dokładniejszych i bardziej użytecznych informacji, co może znacząco wpłynąć na efektywność treningów i zrozumienie gier sportowych na wszystkich poziomach rywalizacji.

3.1.1. Piłka nożna

W problemie analizy piłki nożnej, ostatnie lata przyniosły znaczny postęp, szczególnie w zakresie stosowania technologii wizyjnych [20,61,91,101,107]. Rozwój tych systemów skupia się głównie na trzech aspektach: podsumowaniu zarejestrowanego wydarzenia, dostarczaniu dodatkowych informacji do nagrania oraz analizie wysokiego poziomu. Techniki wizyjne, dostosowane do specyficznych wyzwań środowiska piłkarskiego, umożliwiają ekstrakcję różnorodnych semantycznych poziomów interpretacji, które następnie pomagają sędziom, trenerom oraz widzom dostarczając wartościowe informacje wydobyte w sposób automatyczny.

Jednym z głównych kierunków jest generowanie podsumowania zarejestrowanego wydarzenia, gdzie głównym zadaniem jest wydobycie najciekawszych fragmentów meczu, takich jak momenty zdobycia goli, akcje kartkowe, rzuty rożne czy karne. Realizacja tego zadania wymaga ekstrakcji cech wizualnych niskiego poziomu oraz modelowania ich ewolucji w trakcie rozgrywki.

Innym ważnym obszarem są systemy dostarczające dodatkowe informacje do zarejestrowanego nagrania, które mają za zadanie zwiększyć wartość transmisji telewizyjnych poprzez automatyczne nakładanie dodatkowych danych i wizualizacji na prezentowany obraz. Wymaga to głębszej semantycznej interpretacji scen, co jest możliwe dzięki bardziej złożonym analizom, takim jak detekcja nazwisk graczy zaangażowanych w akcję, określenie rodzaju zdarzenia (np. faul, gol), czy monitorowanie trasy biegu danego zawodnika.

Analiza wysokiego poziomu obejmuje bardziej złożone oceny, wymagające dokładnej lokalizacji pozycji graczy, analizy trajektorii piłki w modelu 3D oraz rozpoznawania zachowań. Semantyczna interpretacja na tym poziomie jest szczególnie wymagająca, gdy zakłada się minimalną interwencję człowieka oraz konieczność przetwarzania danych w czasie rzeczywistym, co jest kluczowe dla bezpośredniego wykorzystania wyników systemu. Rozwój metod wizyjnych w piłce nożnej ilustruje, jak technologie te mogą przyczyniać się do głębszego zrozumienia gry, oferując narzędzia wspomagające zarówno trenerów, jak i analityków sportowych w ich pracy [20].

Praca [61] opisuje innowacyjne podejście do oceny wydajności piłkarzy przy użyciu wielu kamer, koncentrując się na analizie ewolucji postawy ciała zawodników. Autorzy opracowali metodę ekstrakcji cech wyglądu ciała i wykorzystali najbardziej znaczące z nich do modelowania aktywności graczy podczas gry. Zastosowanie ciągłych ukrytych modeli Markowa do modelowania czasowej ewolucji cech ciała w podejściu opartym na wielu kamerach, pozwala na efektywną analizę zachowań graczy, co zostało przetestowane na sekwencjach zachowań graczy z włoskiej ligi Serie A.

W badaniu podkreślono, że istotne jest nie tylko mierzenie pokonanych odległości przez

graczy, ale także ocena, jak długo gracze aktywnie uczestniczyli w grze. Do tego celu analiza postawy ciała jest kluczowa, aby odróżnić różne aktywności graczy, takie jak kontrola piłki czy interakcje z przeciwnikiem prowadzącym grę. Proponowane podejście składa się z dwóch faz: fazy uczenia się, w której pewne sekwencje aktywności graczy są ekstrahowane, oceniane i stosowane do generowania ciągłych modeli Markowa, oraz fazy testowania, w której kolejne przesuwne okna aktywności graczy są poddawane ocenie przy użyciu modeli Markowa.

Eksperymenty przeprowadzone na różnych sekwencjach aktywności graczy wykazują potencjał zastosowania technologii wizyjnych w rzeczywistych scenariuszach sportowych. System wielokamerowy, rozmieszczony po obu stronach boiska, zapewniał pokrycie każdej strefy gry przez dwie przeciwległe kamery, co znacząco zwiększało dokładność analizy. Każdy węzeł przetwarzający obrazy z kamer analizował sekwencje obrazów niezależnie, podczas gdy węzeł nadzorujący zbierał wszystkie dane i dokonywał końcowej oceny prawdopodobieństwa zachowania.

Podsumowując, opracowany przez autorów system wielokamerowy do oceny wydajności piłkarzy na podstawie analizy ewolucji postawy ciała stanowi ważny krok w automatycznej analizie wideo w sporcie. Dalsze prace nad systemem mogą obejmować bardziej szczegółową automatyczną analizę wydajności graczy poprzez wprowadzenie wielu modeli Markowa do modelowania różnych aktywności, takich jak bieganie, uderzanie piłki, odbiór piłki i tym podobne, co może przyczynić się do głębszego zrozumienia gry i poprawy wyników sportowych.

3.1.2. Piłka ręczna

W pracach [12, 89], autorzy skupiają się na wyzwaniach i rozwiązaniach związanych z detekcją obiektów w scenach sportowych, konkretnie w piłce ręcznej, co ma znaczenie również dla innych dyscyplin, w tym piłki nożnej. Kluczowym elementem jest zastosowanie konwolucyjnych sieci neuronowych (ang. convolution neural networks, CNN) do detekcji obiektów, takich jak gracze i piłka, w zmiennych warunkach oświetleniowych i w scenach z częściowymi zasłonięciami tych obiektów. Badanie koncentruje się na porównaniu wydajności różnych obecnych detektorów, w tym YOLO (ang. you only look once), Mask R-CNN (ang. region based convolutional neural network), i mieszanki modeli gaussowskich (ang. gaussian mixture models, GMM), na niestandardowym zbiorze danych zawierającym nagrania z treningów i meczów piłki ręcznej.

Systemy te muszą radzić sobie z dynamicznym środowiskiem sportowym, gdzie zmienne warunki oświetleniowe, szybki ruch graczy, i różnorodność rozmiarów obiektów stanowią istotne wyzwanie. Podejścia bazujące na modelach YOLO i Mask R-CNN, stosując głębokie sieci neuronowe, oferują nowe możliwości w detekcji i klasyfikacji obiektów, jednak ich skuteczność może być ograniczona przez szybkość przetwarzania i wymagania dotyczące mocy obliczeniowej. Z kolei GMM, stosując podejście do odejmowania tła, zapewnia szybką identyfikację ruchomych obiektów, ale może być mniej skutecznym rozwiązaniem w skomplikowanych scenach.

Porównanie wydajności tych detektorów na scenach z piłki ręcznej pokazuje ich mocne i słabe strony w różnych scenariuszach, co ma bezpośrednie przełożenie na ich potencjalne zastosowanie w analizie wideo innych sportów np. piłki nożnej. Model YOLO, ze względu na swoją szybkość, wydaje się być dobrym wyborem dla analiz w czasie rzeczywistym, podczas gdy Mask R-CNN, zapewniający dokładniejszą segmentację obiektów, może być bardziej odpowiedni dla szczegółowych analiz po meczu. GMM, mimo swojej prostoty, może służyć

jako narzędzie do szybkiego wyodrębniania ruchomych obiektów, ale z ograniczoną zdolnością do ich klasyfikacji.

Wnioski z prac podkreślają znaczenie doboru odpowiedniego narzędzia w zależności od specyficznych wymagań analizy sportowej, równoważąc między dokładnością, a szybkością przetwarzania. Rozwój i optymalizacja tych systemów detekcji obiektów mają potencjał do znacznego wzbogacenia analizy taktycznej i technicznej w sporcie, oferując trenerom i analitykom nowe narzędzia do oceny wydajności graczy i zespołów.

3.1.3. Tenis

Badanie przedstawione w pracy [102] koncentruje się na automatycznej analizie nagrań wideo z meczów tenisa. Autorzy proponują metodę, która umożliwia automatyczną detekcję linii kortu tenisowego oraz śledzenie graczy na sekwencjach klatek wideo. Dodatkowo, przedstawiają algorytm oparty na kolorze do selekcji fragmentów nagrań z kortem tenisowym z surowego materiału wideo, co jest kluczowe dla dalszej analizy.

Celem pracy jest automatyzacja generowania użytecznych i wysokopoziomowych oznaczeń, takich jak rodzaje wymian (np. gra z linii końcowej, zagrania przechodzące, gra przy siatce) dla segmentów wideo. Takie oznaczenia mają na celu ułatwienie profesjonalnym tenisistom i trenerom przeszukiwanie posiadanych nagrań wideo w bardziej efektywny sposób.

Do realizacji celu, badacze opracowali dwuetapowy system, gdzie w pierwszym etapie selekcjonowane są fragmenty wideo zawierające tylko korty tenisowe za pomocą algorytmu opartego na kolorze. Następnie, na wyselekcjonowanych fragmentach, wykrywane są linie kortu i śledzeni są gracze za pomocą opracowanego algorytmu, co umożliwia dalszą wysokopoziomową analizę i klasyfikację wydarzeń w grze.

Eksperymentalne wyniki na rzeczywistych danych wideo z tenisa pokazują ważność i skuteczność podejścia. Praca ta podkreśla znaczenie odpowiedniego wyboru narzędzi w zależności od specyficznych wymagań analizy sportowej, balansując między dokładnością a szybkością przetwarzania. Rozwój i optymalizacja systemów detekcji obiektów ma potencjał znaczącego wzbogacenia analizy taktycznej i technicznej w sporcie, oferując nowe narzędzia do oceny wydajności graczy i drużyn [102].

3.1.4. Koszykówka

Analiza zawarta w pracy [40] koncentruje się na wykorzystaniu technik wizji komputerowej do zbierania danych statystycznych w sporcie, ze szczególnym naciskiem na koszykówkę. Autor pracy przedstawia metodologię tworzenia rozwiązania zdolnego do automatycznego zbierania danych o pozycjach graczy i piłki. Dzięki zastosowaniu zaawansowanych algorytmów wizji komputerowej, praca ma na celu ułatwienie analizy gry zarówno dla zespołów profesjonalnych, jak i szkół średnich czy uczelni wyższych, które nie mogą sobie pozwolić na drogie systemy komercyjne.

System zaprojektowany przez autora składa się z kilku kluczowych elementów, w tym zestawu kamer do zbierania obrazów z gry oraz oprogramowania do analizy tych obrazów w celu identyfikacji pozycji graczy i piłki. Przy wykorzystaniu dostępnych bibliotek, autor opracował metodę do automatycznego śledzenia ruchów graczy i piłki, co umożliwia zbieranie danych o ich pozycjach w czasie rzeczywistym.

Poza systemem do zbierania danych statystycznych, autor bada również możliwość analizy rzutów wolnych w koszykówce, badając, czy istnieje specyficzna relacja między formą, a wynikiem rzutu. Wykorzystując techniki wizji komputerowej do analizy postawy ciała zawodników podczas rzutów, system ma na celu dostarczenie informacji zwrotnej, która może pomóc w poprawie techniki zawodników.

Podsumowując, praca ta prezentuje kompleksowe podejście do analizy sportowej za pomocą wizji komputerowej, pokazując, jak technologie wizyjne mogą być zastosowane do zbierania i analizy danych sportowych. Dzięki temu zespoły na różnych poziomach rywalizacji mają możliwość uzyskania dostępu do cennych informacji statystycznych, które mogą przyczynić się do poprawy strategii gry i wydajności zawodników [40].

3.2. Dostępne bazy danych

Niniejszy podrozdział zawiera przegląd dostępnych baz danych, które stanowią kluczowe zasoby dla rozwoju i ewaluacji algorytmów przetwarzania obrazu oraz uczenia maszynowego w kontekście sportu. Ze względu na unikalne wyzwania stawiane przez dynamiczne i często nieprzewidywalne środowisko sportowe, odpowiednio zaprojektowane i zgromadzone bazy danych odgrywają fundamentalną rolę w postępie naukowym w dziedzinie wizji komputerowej. Szczególnie w sporcie, gdzie szybkie ruchy, złożone interakcje między zawodnikami, a także specyficzne dla danej dyscypliny reguły i sytuacje świadczą o wysokim stopniu skomplikowania analizy obrazu. Potrzebne są zatem zasoby danych umożliwiające skuteczną pracę nad nowymi metodami klasyfikacji, detekcji i śledzenia obiektów.

Kluczowym aspektem, na który zwraca się uwagę przy analizie dostępnych zbiorów danych, jest ich różnorodność i reprezentatywność. W dziedzinie sportu, gdzie warunki oświetleniowe, konfiguracje kamery i typy aktywności mogą znacząco się różnić, bazy danych muszą odzwierciedlać te zmienności, aby algorytmy były w stanie skutecznie radzić sobie w różnorodnych scenariuszach. Ponadto, jakość danych, ich rozdzielczość, a także sposób oznaczania informacji są kluczowe dla precyzji i użyteczności modeli uczenia maszynowego.

Omawiając poszczególne bazy danych, skoncentrowano się na charakterystyce technicznej nagranych materiałów, takich jak długość filmów, rozdzielczość i częstotliwość klatek na sekundę (ang. frames per second, fps), które są istotne z punktu widzenia przetwarzania wideo i analizy ruchu. W kontekście sportu, gdzie szybkość i płynność ruchów odgrywają kluczową rolę, parametry te mają bezpośredni wpływ na możliwości detekcji i śledzenia obiektów, a tym samym na skuteczność budowanych rozwiązań.

Przeanalizowano również sposób, w jaki dane zostały oznaczone i jakie informacje zawierają etykiety. Precyzyjne oznaczanie akcji sportowych, pozycji zawodników, a także specyficznych elementów obrazu, takich jak piłka czy inne obiekty, jest niezbędne dla rozwoju algorytmów zdolnych do interpretacji złożonych scen sportowych. Podsumowując, ten podrozdział ma na celu nie tylko przedstawienie obecnego stanu dostępnych baz danych, ale także podkreślenie ich znaczenia dla badań i rozwoju technologii wizji komputerowej w sporcie.

W obecnych otwartych repozytoriach znajdują się między innymi następujące bazy danych [108]:

- **Sports Videos in the Wild** - zbiór danych składa się z 4 200 nagrań wideo zarejestrowanych za pomocą smartfonów. Baza obejmuje 30 kategorii sportowych i 44 różne akcje.

Każde wideo jest oznaczone gatunkiem sportu, a dla 40% wideo określono również ramy czasowe trwania każdej akcji.

- **ISSIA Soccer** - zbiór danych zawiera wideo z meczu piłki nożnej, zarejestrowane za pomocą 6 kamer (rozdzielczość 1920x1088 px w 25 fps) umieszczonych po obu stronach boiska. Jest to kompleksowy zbiór, z manualnymi adnotacjami pozycji graczy, sędziów i piłki w każdej klatce, wraz z obrazami kalibracyjnymi dla każdej kamery. Zbiór ten jest szczególnie użyteczny dla badań skupiających się na śledzeniu graczy i piłki w piłce nożnej, oferując materiały wysokiej jakości z różnych perspektyw.

- **UIUC2** - zbiór danych zawiera sekwencje nagrań wideo z mistrzostw świata w badmintonie z 2006 roku, obejmujące zarówno mecze w pojedynkę, jak i w parach. Każda sekwencja została szczegółowo oznaczona pod kątem typów ruchów, typów uderzeń, i momentów uderzenia piłki, a także zawiera informacje o lokalizacji graczy i masce pierwszego planu. Jest to szczególnie przydatne dla badań nad śledzeniem ruchów sportowców i analizą ich technik w grze.

- **APIDIS Basketball** - baza danych zawiera 16 minut nagrania meczu koszykówki z 7 różnych kamer rozmieszczonych wokół i nad boiskiem koszykarskim, co oferuje wszechstronne monitorowanie przebiegu gry. Kamery rejestrowały obraz w rozdzielczości 1600x1200 px w 22 fps. Następnie dane z całego meczu zostały ręcznie oznaczone pod kątem różnych zdarzeń koszykarskich zachodzących w trakcie gry. Pozycje graczy, sędziów, koszy i piłki zostały oznaczone dla jednej minuty. Zestaw zawiera również pomiary i obrazy kalibracyjne umożliwiające kalibrację każdej kamery do wspólnego systemu współrzędnych, co czyni bazę niezwykle wartościową dla badań skoncentrowanych na analizie stosującej wiele kamer jednocześnie w celu wykrywania zdarzeń.

- **Sports 1M** - baza zawiera 1 133 158 adresów url do nagrań wideo, które zostały automatycznie oznaczone stosując 487 etykiet sportowych. Klasy są dodatkowo podzielone w między innymi takie grupy jak: sporty wodne, sporty zespołowe, sporty zimowe, sporty z piłką, sporty walki i sporty z użyciem zwierząt.

- **Leeds Sports Pose** - baza danych składa się z 2 000 obrazów sportowców uprawiających różne dyscypliny sportowe, z adnotacjami pozycji 14 stawów do szacowania pozycji ciała. Baza może być stosowana do badań w dziedzinie wykrywania pozycji i rozpoznawania aktywności w różnorodnych kontekstach sportowych, takich jak lekkoatletyka, badminton, baseball, gimnastyka, piłka nożna, tenis czy siatkówka.

- **UIUC Sports Event** - zbiór danych zawiera zdjęcia z ośmiu różnych kategorii wydarzeń sportowych, takich jak wioślarstwo, badminton, polo, bocce, snowboard, krykiet, żeglarstwo, czy wspinaczka skalna. Zbiór zawiera zdjęcia sportowców podczas wykonywania aktywności, każde z nich jest oznaczone etykietą odpowiadającą konkretnej dyscyplinie sportowej. Zdjęcia te są użyteczne dla badań nad rozpoznawaniem i klasyfikacją aktywności sportowych na podstawie analizy postaw i ruchów sportowców.

- **Olympic Sports** - zbiór oferuje nagrania wideo sportowców trenujących 16 różnych dyscyplin olimpijskich. Zapewnia 50 sekwencji wideo dla każdej aktywności, oznaczone

etykietą klasy działania, co może być niezbędne do szkolenia i oceny modeli klasyfikacji i rozpoznawania aktywności w sporcie.

- **Volleyball Activity** - zbiór danych składa się z nagrań wideo z zawodów profesjonalnej austriackiej ligi siatkówki. Dane obejmują wysokiej jakości (rozdzielczość 1920x1080 px w 25 fps) nagrania meczów, na których oznaczono aktywności poszczególnych graczy, takie jak serwowanie, przyjęcie, atak, blok, czy obrona. Dane te są przeznaczone dla badań nad analizą taktyczną i techniczną w meczach siatkówki.

- **UCF Sports Action Dataset** - zbiór składa się z zestawu akcji zebranych z różnych sportów, które są prezentowane w kanałach telewizji sportowej, takich jak BBC, czy ESPN. Zbiór danych zawiera 10 aktywności sportowych, takich jak nurkowanie, wymach golfowy, kopanie piłki, podnoszenie ciężarów, jazda konna, bieganie, jazda na deskorolce czy chodzenie. Zbiór danych zawiera łącznie 150 sekwencji w rozdzielczości 720x480 px w 10 fps. Dostępne adnotacje to lokalizacja akcji i etykiety klas dla każdej aktywności.

- **CVBASE '06** - baza jest podzielona na trzy części, każda z nich służy różnym celom związanym ze śledzeniem i rozpoznawaniem aktywności w sportach zespołowych. Pierwszy podzbiór obejmuje 10 minutowe nagranie z gry w piłkę ręczną, zarejestrowane z trzech zsynchronizowanych kamer, z dostępnymi trajektoriami dla 7 graczy. Dodatkowo zawiera manualne adnotacje ekspertów sportowych, w tym aktywności drużynowe i indywidualne. Drugi podzbiór zawiera nagrania z dwóch meczów squasha, dostarczając trajektorie i ręczne adnotacje takie jak fazy i rodzaje uderzeń. Trzeci podzbiór składa się z dwóch zsynchronizowanych kamer umieszczonych nad boiskiem, które rejestrują 5 minut gry w koszykówkę. Wszystkie trzy części były rejestrowane kamerami o takich samych parametrach tj. rozdzielczości 384x288 px w 25 fps.

- **S-HOCK** - zbiór danych zawiera nagranie publiczności podczas czterech meczów hokejowych. Podczas rejestrowania zastosowano 5 kamer: 2 szerokokątne kamery HD oraz trzy kamery nakierowane na różne sekcje widowni. Dostępne adnotacje obejmują lokalizację ciała i głowy obserwatorów, ich postawę i akcje. W sumie oznaczono 13 965 klatek i 1 950 210 postaci.

3.3. Analiza zachowań zawodników w boksie olimpijskim

Analiza zachowań zawodników w boksie olimpijskim staje się coraz bardziej istotna w kontekście rosnących wymagań dotyczących optymalizacji treningu i taktyki zawodników. Boks olimpijski, jako dyscyplina sportowa łącząca wymagania fizyczne, taktyczne i psychologiczne, wymaga dogłębnej analizy i zrozumienia zarówno indywidualnych umiejętności zawodników, jak i dynamiki walki. Rozwój technologii i metod uczenia maszynowego oraz przetwarzania obrazu otwiera nowe możliwości dla naukowców i trenerów, umożliwiając bardziej szczegółową i zautomatyzowaną analizę zachowań sportowych.

Współczesne badania skupiają się na różnorodnych metodach analizy, stosujących zarówno techniki inwazyjne, oparte na ubieralnych urządzeniach i sensorach, jak i nieinwazyjne, korzystające z danych wizyjnych. Metody nieinwazyjne zyskują na znaczeniu, zwłaszcza w kontekście

boksu olimpijskiego, gdzie regulacje dotyczące ubioru i wyposażenia zawodników ograniczają możliwości stosowania niektórych metod opartych na sensorach. Skuteczność treningu i przygotowań taktycznych zawodników może być znacząco zwiększona dzięki precyzyjnej analizie ich zachowań w ringu.

Niniejszy podrozdział zawiera przegląd zasad boksu olimpijskiego, z naciskiem na regulacje dotyczące oceny, ubioru, charakterystyki ringu oraz bezpieczeństwa zawodników. Następnie, omawiane są aktualne podejścia do analizy zachowań bokserów, zarówno z perspektywy metod inwazyjnych, jak i nieinwazyjnych, podkreślając ich znaczenie dla poprawy treningów i strategii walki. Ostatnia część podrozdziału koncentruje się na metodologii analizy zachowań bokserów przyjętej w rozprawie. Podrozdział ten ma na celu nie tylko przedstawienie stanu wiedzy, ale również zwrócenie uwagi na potencjał obecnych technologii w analizie i optymalizacji przygotowań zawodników do zawodów bokserskich.

3.3.1. Zasady boksu olimpijskiego

Podrozdział ten ma na celu szczegółową analizę regulacji obowiązujących w boksie olimpijskim, z naciskiem na aspekty takie jak zasady oceny, wymogi dotyczące ubioru, charakterystykę ringu bokserskiego oraz inne istotne kwestie. Zasady boksu są regulowane między innymi przez amerykańskie stowarzyszenie komisji bokserskich (ang. Association of Boxing Commissions, ABC), organizację non-profit, która zapewnia ramy dla organizowania walk bokserskich oraz MMA. W kontekście punktowania walk, zasady ABC określają, że system 10 punktów (ang. 10 Point Must System) jest standardowym systemem, który jest najbardziej rozpoznawalnym i stosowanym systemem punktacji od połowy XX wieku [1].

W kontekście boksu olimpijskiego, szczególnie ważne jest podkreślenie faktu, że wszyscy licencjonowani sędziowie zaangażowani w przeprowadzenie wydarzenia podlegają bezpośredniej kontroli komisji nadzorującej wyznaczonej do regulowania wydarzenia. Istotnym aspektem jest, że sędziowie nie mogą wykazywać stronniczości względem żadnego z zawodników. Każdy bokser przed walką poddawany jest obowiązkowym badaniom fizycznym przez lekarza przy ringu, który pisemnie certyfikuje zdolność fizyczną zawodnika do bezpiecznego udziału w zawodach.

Regulacje szczegółowo określają dopuszczalny ubiór i sprzęt ochronny dla zawodników, którzy powinni rywalizować w spodenkach bokserskich, ochraniaczu brzucha, ochraniaczu krocza, butach i indywidualnie dopasowanym ochraniaczu na zęby. Ponadto kobiety powinny dodatkowo nosić koszulkę na ciało, podczas gdy ochraniacze piersi są opcjonalne.

Procedury ważenia

Ważenie zawodników ma kluczowe znaczenie dla zapewnienia uczciwości i bezpieczeństwa zawodów. Musi ono odbyć się w ciągu 24 godzin przed zaplanowanym wydarzeniem, w obecności komisji nadzorującej i przedstawiciela promotora. Procedury ważenia obejmują również szczegółowe zasady dotyczące dopuszczalnych różnic w wadze między zawodnikami, zależnie od kategorii wagowej.

Ring bokserski i bezpieczeństwo

Wymiary ringu, jak również wymagany sprzęt medyczny i personel, są ściśle określone w celu

[1] https://www.abcboxing.com/abc-regulatory-guidelines/ (Data dostępu: 04.01.2021)

zapewnienia bezpieczeństwa zawodników. Ring musi mieć co najmniej 16 stóp kwadratowych powierzchni wewnętrznej i być wyposażony w cztery liny. Ambulans wraz z licencjonowanymi technikami medycznymi ratunkowymi musi być obecny przy ringu przez cały czas trwania zawodów.

Kary i faule
Komisja ABC dostarcza także szczegółowe wytyczne dotyczące fauli i odpowiednich kar, w tym okoliczności, w których zawodnik może zostać zdyskwalifikowany lub walka może zostać zatrzymana z powodu nieprzestrzegania zasad. To obejmuje nie tylko nielegalne uderzenia, ale również zachowania, takie jak celowe wypuszczanie ochraniacza na zęby, uderzenie przeciwnika poniżej pasa, czy użycie nieczystych technik walki.

Analiza tych regulacji pozwala na głębsze zrozumienie zasad panujących w boksie olimpijskim, podkreślając zarówno znaczenie bezpieczeństwa zawodników, jak i ducha sportowej rywalizacji. Przepisy te są fundamentem, na którym budowana jest struktura zawodów bokserskich, zapewniając, że wszystkie aspekty zawodów - od ważenia, poprzez sprzęt, aż po zachowanie w ringu - są ściśle regulowane w celu zapewnienia uczciwej i bezpiecznej rywalizacji.

Owijanie dłoni
Owijanie dłoni jest kluczowym aspektem przygotowań boksera do walki, mającym na celu ochronę dłoni i nadgarstków. Zgodnie z regulaminem, owijki na dłonie muszą być ograniczone do maksymalnie dwudziestu jardów miękkiej gazy, nie szerszej niż dwa cale. Gaza musi być utrzymywana na miejscu przez nie więcej niż osiem stóp taśmy klejącej, nie szerszej niż jeden i pół cala. Taśma klejąca nie może pokrywać żadnej części kostek (kłykci) dłoni, gdy ręka jest zaciśnięta w pięść. Stosowanie wody lub jakichkolwiek innych płynów lub materiałów na taśmie jest surowo zabronione.

Owijki muszą być aplikowane w szatni w obecności przedstawiciela komisji i jednego przedstawiciela drugiego boksera, jeśli sam o to poprosi. Ścisłe regulowanie procesu owijania dłoni ma na celu zapewnienie równych warunków dla obu zawodników oraz ochronę zdrowia i bezpieczeństwa sportowców.

Rękawice
Rękawice bokserskie są podstawowym elementem wyposażenia każdego zawodnika, mającym za zadanie ochronę zarówno uderzającego, jak i odbierającego cios. Każdy zestaw rękawic musi posiadać część kciuka przyczepioną do korpusu rękawicy, co ma na celu minimalizację ryzyka kontuzji oka przeciwnika. Rękawice, lub zestaw rękawic, może być używany tylko raz podczas każdego wydarzenia bokserskiego. Wszystkie rękawice podlegają inspekcji przez nadzorującą komisję. Rękawice uznane za zniekształcone, zmodyfikowane, nieodpowiednie, lub źle dopasowane, muszą zostać wymienione.

Promotor wydarzenia zobowiązany jest dostarczyć komisji przed rozpoczęciem pierwszej walki po jednym komplecie rękawic o grubości 8 i 10 uncji do wykorzystania, na wypadek uszkodzenia rękawic podczas zawodów. Promotorzy muszą zapewnić nowe rękawice na wszystkie główne walki i walki o tytuł.

Osoby w ringu
Podczas walki bokserskiej, w ringu mogą znajdować się wyłącznie zawodnicy oraz sędzia. Dla walk nieoferujących tytułu mistrzowskiego, w narożniku może znajdować się nie więcej niż trzech sekundantów. Pomiędzy rundami jeden sekundant może być wewnątrz ringu, a dwóch na jego apronie (powierzchni ringu bokserskiego rozciągającego się poza linami). Podczas walki o mistrzostwo liczba sekundantów może wzrosnąć do czterech, z czego jeden może znajdować się wewnątrz ringu, a dwóch na apronie, przy czym czwarty sekundant musi pozostać na podłodze, poza apronem.

Lekarz może wejść do ringu, jeśli zostanie poproszony przez sędziego, komisję nadzorującą lub inspektorów, aby zbadać uraz zawodnika. Żaden z zawodników nie może opuścić ringu podczas jednominutowej przerwy między rundami. Sędzia ma prawo zatrzymać walkę, jeśli nieautoryzowana osoba wejdzie na ring podczas rundy.

Nokdaun (ang. knockdown)
Sytuacja, w której zawodnik znajdzie się na podłodze ringu, może zostać uznana za nokdaun, jeśli zostanie spowodowana przez dozwolony cios lub serię ciosów. Do nokdaunu dochodzi, gdy zawodnik:

- dotknie podłogi dowolną częścią ciała poza stopami.
- Jest podtrzymywany przez liny lub zawisł na linach, przez lub nad nimi, nie będąc w stanie się obronić i nie mogąc spaść na podłogę.

Liczenie do ośmiu po nokdaunie
Po każdym nokdaunie obowiązuje liczenie do ośmiu. Sędzia może zakończyć liczenie i walkę w dowolnym momencie, gdy uzna, że bezpieczeństwo leżącego zawodnika jest zagrożone. Jeśli zawodnik podniesie się przed osiągnięciem liczenia i natychmiast ponownie upadnie na podłogę bez otrzymania ciosu od przeciwnika, sędzia wznowi liczenie w miejscu, w którym je przerwał.

W momencie nokdaunu przeciwnik zawodnika leżącego na podłodze musi udać się do najdalszego neutralnego narożnika i pozostać tam podczas liczenia. Sędzia może przerwać liczenie, jeśli przeciwnik nie uda się do neutralnego narożnika, i wznowić liczenie od momentu przerwania, gdy przeciwnik znajdzie się w wymaganym miejscu.

Zawodnik, który zostaje znokautowany, musi zostać zawieszony na minimalny okres sześćdziesięciu dni, a zawodnik, który przegra przez techniczny nokaut, musi zostać zawieszony na minimalny okres trzydziestu dni od udziału w jakiejkolwiek aktywności bokserskiej.

Obowiązki sędziego przed walką
Przed rozpoczęciem walki, sędzia ma szereg obowiązków, które musi spełnić w celu zapewnienia bezpieczeństwa i uczciwości zawodów. Do tych obowiązków należą:

1. Spotkanie z każdym z bokserów i ich głównym sekundantem w szatni, aby:
    - poinformować o zasadach dotyczących sekundantów i konsekwencjach ich nieprzestrzegania.
    - Zaznaczyć linie pasa/bioder, wyjaśniając, że sprzęt nie może wystawać powyżej tej linii.

- Przekazać informacje o procedurach końca rundy.
- Omówić zarządzanie narożnikiem przez głównego sekundanta.
- Przedyskutować procedury przerwania walki i zachowania w przypadku fauli.
- Sprawdzić i podpisać owijki na dłonie.

2. Spotkanie z lekarzem przy ringu, aby:
   - omówić doświadczenie lekarza w pracy przy ringu.
   - Przekazać, jakie sygnały będą używane do wezwania lekarza do ringu.
   - Upewnić się, że lekarz nie będzie wchodził do ringu bez polecenia sędziego.
3. Sprawdzenie stanu ringu, aby upewnić się, że wszystko jest gotowe na walkę.
4. Przeprowadzenie końcowego sprawdzenia z bokserami w ringu, w tym:
   - inspekcja rękawic, sprzętu i wyglądu zawodników.
   - Przywołanie bokserów do środka ringu, udzielenie ostatnich instrukcji i sprawdzenie gotowości sędziów punktowych.

Obowiązki sędziego podczas walki
Podczas walki, sędzia ma za zadanie:

1. zapewnić bezpieczeństwo bokserów.
2. Egzekwować przestrzeganie zasad.
3. Utrzymywać kontrolę nad walką, w tym wydawanie ostrzeżeń i ewentualne odejmowanie punktów.
4. W przypadku dotknięcia rękawic zawodnika z podłogą (czy to przez przypadek, czy w wyniku przewrócenia), przeprowadzić inspekcję rękawic i wytrzeć je do czysta przed kontynuacją walki.
5. W przypadku obrażeń, może przerwać walkę, aby skonsultować się z lekarzem ringowym.
6. Zarządzać sytuacjami spornymi, w tym decydować o przerwaniu walki z powodu faulu lub kontuzji.

Obowiązki sędziego po zakończeniu walki
Po zakończeniu walki, sędzia:

1. zbiera karty punktowe od sędziów i przekazuje je komisji.
2. Sprawdza owijki na dłonie zawodników po zdjęciu rękawic.
3. Przywołuje bokserów do środka ringu i ogłasza zwycięzcę.

4. Utrzymuje porządek w ringu i na jego obrzeżach do momentu opuszczenia go przez bokserów i sekundantów.

Te procedury zapewniają, że walki bokserskie są prowadzone w sposób sprawiedliwy, bezpieczny i zgodny z ustalonymi zasadami. Sędzia pełni kluczową rolę w zarządzaniu przebiegiem walki, dbając o przestrzeganie regulaminu i ochronę zdrowia zawodników.

Zasady oceniania

W boksie, ocena walki i przyznawanie punktów odbywa się przez specjalnie wyznaczonych sędziów, którzy są zatwierdzeni przez komisję nadzorującą walkę. Rolą sędziów jest ocenianie każdej rundy walki niezależnie, stosując ustalone kryteria oceny. Sędzia ringowy (ang. referee), mimo pełnienia kluczowej roli w zarządzaniu przebiegiem walki, nie uczestniczy w ocenianiu walki ani przyznawaniu punktów.

Oceny dokonuje trzech sędziów punktowych, których wybór zatwierdza nadzorująca komisja. Dzięki temu zapewniona jest obiektywna i zróżnicowana perspektywa na przebieg pojedynku.

Ocenianie w boksie odbywa się na podstawie systemu 10 punktów (ang. 10 Point Must System), co oznacza, że zwycięzca rundy otrzymuje 10 punktów, a przegrany mniej – zazwyczaj 9, a w przypadku nokdaunów lub zdominowania rundy punkty mogą być odjęte w większej liczbie. Kryteria, według których sędziowie przyznają punkty, obejmują:

1. czyste uderzenia (ang. clean punching) - mocne i skuteczne ciosy mają większą wartość niż duża liczba słabych trafień.

2. Efektywna agresja (ang. effective aggressiveness) - zawodnik musi być agresywny, ale również efektywnie trafiać przeciwnika, a nie tylko atakować bez skutku.

3. Ogólna kontrola ringu (ang. ring generalship) - ogólna zdolność kontrolowania przebiegu walki przez zawodnika.

4. Obrona (ang. defense) - umiejętność unikania ciosów przeciwnika oraz obrony przed atakami.

Odejmowanie punktów

Sędziowie odejmują punkty za nokdauny, jak również mogą odejmować punkty za faule, gdy zostaną do tego poinstruowani przez sędziego ringowego. Ważne jest, aby sędziowie stosowali te zasady konsekwentnie, aby zapewnić sprawiedliwą ocenę walki.

Podsumowanie

Rola sędziów w boksie jest niezwykle ważna, ponieważ to oni, poprzez obiektywną ocenę według ustalonych kryteriów, decydują o wyniku walk. Ich zadaniem jest nie tylko liczenie ciosów, ale również ocena jakości walki, strategii i umiejętności obronnych zawodników. Proces oceniania wymaga nie tylko doskonałej znajomości zasad boksu, ale również umiejętności szybkiego analizowania i interpretowania akcji w ringu.

3.3.2. Aktualne podejścia do analizy zachowań bokserów

Aktualny postęp w obszarze uczenia maszynowego i przetwarzania obrazu otwiera nowe możliwości analizy i zrozumienia złożonych dyscyplin sportowych, takich jak boks olimpijski. Specyfika tego sportu, charakteryzująca się szybkimi i precyzyjnymi ruchami, wymaga zastosowania zaawansowanych metod analitycznych, które pozwolą na dogłębne zbadanie technik, strategii oraz ruchów bokserów. Niniejszy podrozdział zawiera przegląd aktualnych podejść do analizy zachowań zawodników w boksie olimpijskim, z zastosowaniem narzędzi sztucznej inteligencji i przetwarzania obrazu.

Aktualny stan wiedzy wskazuje, że analiza zachowań bokserów może być przeprowadzana za pomocą dwóch głównych grup metod: techniki stosujące sensory i urządzenia ubierane przez zawodników oraz techniki nieinwazyjne oparte na analizie obrazu z kamer. Pierwsze z nich korzystają z urządzeń elektronicznych i sensorów, które rejestrują różnorodne parametry, takie jak ruchy ciała, przyspieszenie, czy siła uderzeń, dostarczając cennych danych na temat fizycznych aspektów walki. Z kolei podejścia nieinwazyjne skupiają się na analizie wizualnej, gdzie zaawansowane algorytmy przetwarzania obrazu i uczenia maszynowego pozwalają na automatyczne rozpoznawanie i ocenę ruchów, postaw oraz technik bokserów, bez konieczności stosowania dodatkowego sprzętu przez sportowców. Obie metody dostarczają istotnych informacji, które mogą być stosowane do poprawy treningów, strategii walki oraz w ocenie kondycji i technik zawodników, podnosząc tym samym poziom przygotowań do zawodów.

Podejścia inwazyjne

Autorzy pracy [118] koncentrują się na analizie wydajności różnych konfiguracji czujników inercyjnych oraz modeli uczenia maszynowego w kontekście klasyfikacji typów ciosów w boksie. Przy użyciu dwóch prostych konfiguracji sensorów – jednej stosującej czujniki umieszczone na nadgarstkach zawodnika, a drugiej dodatkowo wyposażonej w czujnik na trzecim kręgu piersiowym (T3) – autorzy dążą do zautomatyzowania procesu klasyfikacji ruchów, co ma potencjalnie umożliwić pominięcie procesu manualnego etykietowania wideo i zbudować fundamenty pod automatyczne monitorowanie obciążenia treningowego sportowców uprawiających sporty walki.

Do zbierania danych zastosowano 9 stopniowe czujniki swobody, które rejestrowały ruchy zawodników podczas wykonywania różnych ciosów. Autorzy przeprowadzili trening i testowanie sześciu modeli uczenia maszynowego, w tym regresji logistycznej, maszyny wektorów nośnych (ang. support vector machines, SVM) i wielowarstwowego perceptronu (ang. multilayer perceptron, MLP), zarówno w wersjach dostrojonych, jak i niedostrojonych.

Rezultatem badań była wysoka skuteczność klasyfikacji typów uderzeń za pomocą obu konfiguracji sensorów, z dokładnością wynoszącą średnio $0,90$ dla sensorów umieszczonych na nadgarstkach i $0,87$ dla konfiguracji z dodatkowym czujnikiem na kręgu T3. Te wyniki demonstrują potencjalne zastosowanie ubieralnych technologii sensorowych i zaawansowanych modeli uczenia maszynowego w analizie i optymalizacji treningu bokserskiego.

Badanie to podkreśla możliwości automatyzacji analizy zachowań sportowców w sportach walki, oferując nowe narzędzia do monitorowania treningu i poprawy wydajności zawodników. Praca pokazuje kierunki dalszych badań w dziedzinie wykorzystania technologii ubieralnych i uczenia maszynowego w sporcie, z potencjalnym wpływem na trening i ocenę wyników

w boksie [118].

W pracy [124] autorzy przedstawiają nowatorskie podejście do monitorowania siły i prędkości uderzeń w boksie, stosując ubieralne czujniki tekstylne. Metodologia badań opiera się na zastosowaniu elastycznych czujników tekstylnych, zdolnych do wykrywania zarówno dotyku, jak i ruchów bezdotykowych, dzięki zastosowaniu hierarchicznie ułożonej tkaniny przestrzennej jako warstwy dielektrycznej oraz elektrod z tkaniny pokrytej niklem. Czujniki te charakteryzują się wysoką czułością na dotyk oraz niskim limitem wykrywania, co pozwala na dokładne monitorowanie nie tylko siły, ale także prędkości ciosów boksera.

Kluczowym aspektem badania było zastosowanie technologii tekstylnych do tworzenia czujników, które są nie tylko wysoce czułe i precyzyjne, ale także komfortowe w noszeniu i łatwe w integracji z odzieżą treningową. Opracowane czujniki BAT (ang. bimodal all-textile) demonstrują nie tylko doskonałą zdolność do wykrywania zmian w mikrośrodowisku wokół czujnika, takich jak ciśnienie czy odległość, ale także zapewniają komfort noszenia dzięki właściwościom tekstylnym, takim jak przepuszczalność powietrza i wilgoci. Ponadto, zastosowanie prostej i ekonomicznej metody produkcji sprawia, że czujniki te są obiecujące dla szeroko zakrojonych zastosowań w monitorowaniu aktywności fizycznej oraz w inteligentnych systemach treningowych dla sportów walki.

Podsumowując, praca ta przedstawia znaczący postęp w aspekcie tekstyliów i ubieralnych technologii sensorycznych, oferując efektywne rozwiązania dla monitorowania w czasie rzeczywistym siły i prędkości uderzeń w boksie, co ma znaczący wpływ na rozwój inteligentnych systemów treningowych w sportach walki [124].

Autorzy [49] w swoim badaniu koncentrują się na automatyzacji pomiaru prędkości ciosów bokserów z zastosowaniem jednostek pomiaru inercyjnego (ang. inertial measurment units, IMU) opartych na sztucznej sieci neuronowej. Głównym celem było zautomatyzowanie procesu pomiarowego kinematycznych charakterystyk ciosów, co jest kluczowe dla efektywnego rozwoju sportowców, wymagającego stałego monitorowania prędkości ich ciosów. Do eksperymentów użyto IMU zamontowanych na nadgarstkach bokserów, rejestrujących absolutne przyspieszenie i prędkość kątową. Badanie przeprowadzono na trzech grupach bokserów z różnymi poziomami zaawansowania treningowego.

W badaniu zastosowano wielowarstwowy perceptron (MLP) jako model sieci neuronowej, z parametrami wejściowymi obejmującymi absolutne przyspieszenie i prędkość kątową. Opracowany model wykazał wysoki poziom rozpoznawania ciosów dla wszystkich klas (prosty, sierpowy, hak, ruch bez ciosu), co pozwala wnioskować, że użycie IMU i sztucznej sieci neuronowej znacząco przyspiesza zbieranie danych o zadawanych ciosach przez bokserów umożliwiając ich dalszą analizę.

Osiągnięte wyniki demonstrują, że zastosowanie sieci neuronowych może znacząco usprawnić proces zbierania danych na temat kinematycznych charakterystyk ciosów w boksie. Szczególnie istotnym wynikiem prac jest wysoki poziom jakości klasyfikacji typów ciosów, co przedstawia potencjał zastosowania tej metody do monitorowania i analizy treningu w różnych dyscyplinach walki [49].

Podejścia nieinwazyjne

W kontekście boksu olimpijskiego, nieinwazyjne metody analizy, stosujące dane wizyjne z kamer, umożliwiają obserwację i ocenę technik bokserskich bez fizycznej ingerencji w ubiór zawodnika. Dzięki zaawansowanym algorytmom przetwarzania obrazu i uczenia maszynowego,

te podejścia pozwalają na precyzyjne śledzenie ruchów i identyfikację ciosów, oferując trenerom i zawodnikom obiektywne narzędzie do oceny wydajności sportowej i optymalizacji treningu.

W pracy [66] autorzy skupiają się na prognozowaniu ruchów zawodników w boksie za pomocą danych wizyjnych pochodzących z kamer RGB, stosując do tego rekurencyjne sieci neuronowe (ang. recurrent neural network, RNN). Celem jest stworzenie trenera bokserskiego, który na podstawie wizualnych danych wejściowych jest w stanie przewidzieć następne ruchy zawodnika. W pracy badane i porównywane są wydajności sześciu różnych architektur sieci neuronowych.

Do badań zastosowano jedno nagranie walki autora oraz dane z otwartego zbioru YT8M, zawierające różnorodne ujęcia i tła. Zastosowanie sieci RNN z warstwami LSTM (ang. long short-term memory) umożliwia efektywne przetwarzanie sekwencji nagrania, zachowując informacje historyczne. Wyniki eksperymentów wskazują na możliwości struktury LSTM w dokonywaniu wiarygodnych prognoz na podstawie niewielkich danych uczących, a także na potencjalną skuteczność metod w treningu indywidualnym.

Podsumowując, badanie to dostarcza istotnych wniosków na temat możliwości stosowania zaawansowanych technik przetwarzania obrazu i sieci neuronowych do prognozowania ruchów w boksie. Potwierdza potencjał obecnych technologii w przedstawionych zastosowaniach. Przyszłe badania mogą rozszerzyć zakres danych treningowych, aby uwzględnić większą różnorodność ruchów i technik, co może jeszcze bardziej zwiększyć dokładność i uniwersalność opracowanych modeli predykcyjnych [66].

Autorzy w pracy [81] skupili się na zastosowaniu technologii wizji komputerowej do rozpoznawania działań człowieka (ang. human action recognition, HAR) w sportach walki w celu automatycznego klasyfikowania ruchów ludzkich na podstawie obrazu wideo. Celem pracy jest stworzenie rozwiązania do analizowania i oceniania technik wielu sportowców, a także identyfikacji obszarów możliwych do ulepszenia.

Autorzy skupiają się na stworzeniu zaawansowanych systemów wizji komputerowej do analizy i oceny sportów walki, w tym boksu, kickboxingu i MMA, z zastosowaniem technik detekcji i śledzenia obiektów w czasie rzeczywistym oraz modeli rozpoznawania zachowań ludzkich (HAR) do oceny segmentów wideo z zawodów. Kluczowe cele pracy to oszczędność czasu, redukcja błędów ludzkich poprzez automatyzację generacji danych, oraz stworzenie obszernej bazy danych oznaczonych wideo z różnych sportów walki, co umożliwi porównanie różnych strategii i technik.

W ramach metodologii, badanie stosuje różne techniki wizji komputerowej do klasyfikacji i śledzenia podstawowych akcji w czasie rzeczywistym, takich jak uderzenia, kopnięcia i bezczynność. Stosowane są technologie głębokiego uczenia, w tym YOLOv5, do detekcji i śledzenia obiektów, oraz szereg technik augmentacji danych. W badaniu przeprowadzono testy na różnorodnych zestawach danych wideo, w tym olimpijskim boksie, MMA i sesjach treningowych pojedynczej osoby z workiem treningowym, aby ocenić wydajność i dokładność modeli CV w różnych sportach walki.

Wyniki wskazują na możliwość połączenia detekcji i śledzenia obiektów w czasie rzeczywistym z estymacją pozycji do generowania statystyk wydajności na podstawie ruchów sportowców w nagraniach wideo. Eksperymenty pokazują, że jest możliwe dokładne śledzenie i klasyfikacja różnych aktywności sportowców w czasie rzeczywistym. Wyniki sugerują, że opracowany system wizji komputerowej może efektywnie rozpoznawać specyficzne ruchy w sportach walki i generować statystyki w czasie rzeczywistym, co otwiera drogę do dalszych

badań nad ulepszaniem architektur detekcji i śledzenia obiektów, aby poprawić dokładność i wydajność wyników [81].

Badania przedstawione w pracach [45, 46] prezentują zaawansowaną metodologię do automatycznej klasyfikacji ciosów bokserskich, korzystając z obrazu RGB dodatkowo wzbogaconego informacjami o głębi (obraz RGB-D). Metoda stosuje kamery umieszczone nad głową bokserów co jednocześnie przyczynia się do łagodzenia problemów związanych z zasłonięciami. Rozpoznawanie ciosów jest realizowane poprzez klasyfikatory SVM (ang. support vector machine) i lasów losowych (ang. random forest), stosując kombinacje cech. Proponowane podejście było testowane na sekwencji obrazów z boksu nagranych w Australijskim instytucie sportu z udziałem 14 elitarnych bokserów. Wyniki demonstrują efektywność metody rozpoznawania akcji, gdzie klasyfikator SVM osiągnął dokładność na poziomie 97,3%, poprawiając wyniki najnowszych systemów rozpoznawania działań człowieka.

Analiza skupia się na wykorzystaniu danych z kamer RGB-D do precyzyjnej klasyfikacji ciosów bokserskich, wykorzystując do tego zaawansowane algorytmy przetwarzania obrazu. Autorzy badają skuteczność różnych konfiguracji cech i klasyfikatorów, podkreślając potencjał takiego podejścia w treningu sportowym i ocenie wyników zawodników. Metodologia ta umożliwia dalsze badania w dziedzinie automatyzacji analizy zachowań sportowców, oferując cenne narzędzia do monitorowania treningu i poprawy wydajności bokserów bez potrzeby bezpośredniego kontaktu z zawodnikami. Wyniki badania wskazują na znaczący postęp w wykorzystaniu technologii wizyjnych i algorytmów uczenia maszynowego do analizy ruchów w boksie, co może przyczynić się do rozwoju bardziej efektywnych metod treningowych i oceny wydajności w sporcie walki.

3.3.3. Kierunek prac w niniejszej rozprawie

W ostatnich latach rozwój technologii i metod uczenia maszynowego umożliwił znaczące postępy w problemie analizy zachowań sportowców, w tym bokserów. Wśród dostępnych metod analizy, podejścia nieinwazyjne, bazujące na danych wizyjnych zarejestrowanych przez kamery RGB, zdobywają na popularności [47, 49, 81]. Są one szczególnie cenne w boksie olimpijskim, gdzie regulacje dotyczące wyposażenia zawodników ograniczają możliwości zastosowania ubieralnych sensorów [46]. Nieinwazyjne metody analizy, stosujące zaawansowane techniki przetwarzania obrazu i algorytmy uczenia maszynowego, stanowią ważny kierunek w celu automatycznego analizowania bokserów bez konieczności ingerowania w ich ubiór.

Kluczowym wyzwaniem dla efektywnego zastosowania nieinwazyjnych metod analizy jest zebranie i przetworzenie odpowiedniej bazy danych. W tym kontekście, budowa własnej bazy danych zawierającej zróżnicowane sekwencje ruchów bokserów jest niezbędna. Proces ten obejmuje nie tylko pozyskanie danych wizyjnych, ale również ich odpowiednie oznaczenie, co pozwala na skuteczne trenowanie algorytmów klasyfikacji ruchów. W rozprawie szczegółowo opisano etapy tworzenia bazy danych (rozdział 4), od wyboru odpowiednich technik nagrywania, przez proces oznaczania, aż po wstępne przetwarzanie danych, które przygotowało je do etapu analizy (rozdział 5).

W dalszej części rozprawy omówiono zastosowane algorytmy uczenia maszynowego, które umożliwiają klasyfikację scen w boksie olimpijskim na podstawie analizy obrazu. Stosowanie technik uczenia głębokiego, takich jak konwolucyjne sieci neuronowe (CNN) pozwala na efektywne rozpoznawanie subtelnych różnic między poszczególnymi ciosami bokserów. Dokładna

analiza wybranych architektur sieci, ich treningu oraz optymalizacji podkreśla potencjał tych metod w kontekście analizy zachowań zawodników w sporcie.

Ważnym aspektem omówionym w rozprawie jest także ocena skuteczności wybranych metod. Wyniki eksperymentów, w tym porównanie dokładności rozpoznawania poszczególnych ciosów, ilustruje praktyczne zastosowanie nieinwazyjnych technik w treningu i ocenie wydajności bokserów. Analiza błędów i ograniczeń stosowanych metod pozwala na identyfikację kierunków dalszych badań i potencjalnych ulepszeń w automatyzacji analizy zachowań sportowych.

Podsumowując, rozprawa przedstawia kompleksowe podejście do analizy zachowań bokserów w boksie olimpijskim, stosując nieinwazyjne metody oparte na danych wizyjnych. Skupia się na procesie budowy bazy danych, zastosowaniu i ocenie algorytmów uczenia maszynowego, podkreślając potencjał i wyzwania związane z automatyzacją analizy w sporcie. Ta metodyka otwiera nowe możliwości dla trenerów i sportowców do poprawy treningu i taktyki, jednocześnie zachowując zgodność z regulacjami dotyczącymi ubioru zawodników.

4. Przygotowanie danych

Rola danych w algorytmach uczenia maszynowego jest fundamentem, na którym opiera się efektywność modeli uczenia maszynowego (ML). Jako główny składnik tych systemów, dane determinują zdolność algorytmów do nauki, rozpoznawania wzorców oraz dokonywania precyzyjnych predykcji. W dziedzinie przetwarzania i klasyfikacji obrazów, gdzie uczenie nadzorowane stanowi podstawę metodologii, oznaczone dane są niezbędne dla skutecznego funkcjonowania modeli. Każdy obraz w zestawie danych musi posiadać przypisaną mu klasę, dzięki której algorytm jest w stanie nauczyć się, jakie cechy charakteryzują poszczególne kategorie obiektów.

Znaczenie oznaczonych danych w uczeniu nadzorowanym podkreśla potrzebę dysponowania wysokiej jakości zbiorami danych. Taka jakość danych umożliwia precyzyjne trenowanie modeli uczenia maszynowego, co jest kluczowe dla rozwoju technologii rozpoznawania obrazu. Niedokładnie oznaczone lub niewystarczająco zróżnicowane zbiory danych mogą prowadzić do problemów z generalizacją modeli, co z kolei może skutkować niepożądanymi błędami w klasyfikacji nowych, niewidzianych wcześniej przez model obrazów.

Skala danych jest kolejnym kluczowym czynnikiem wpływającym na skuteczność algorytmów uczenia maszynowego. Większe i bardziej zróżnicowane zbiory danych pozwalają modelom na uczenie się i adaptację do bardziej złożonych wzorców, co zwiększa ich zdolność do generalizacji. Dlatego też, w przetwarzaniu obrazu, znaczące jest, aby dysponować obszernymi zbiorami danych, które odzwierciedlają różnorodność rzeczywistego świata.

Proces zbierania i oznaczania danych jest zadaniem wymagającym i czasochłonnym, ale niezbędnym dla tworzenia wartościowych zbiorów danych. Wyzwaniem jest nie tylko znalezienie odpowiedniej ilości danych, ale również ich adekwatne oznaczanie. Wymaga to często wiedzy eksperckiej lub zaawansowanych technik automatycznego etykietowania, które mogą skutecznie identyfikować i klasyfikować obiekty na obrazach.

Dostęp do danych stanowi jedno z głównych wyzwań w dziedzinie sztucznej inteligencji, ze względu na ograniczenia prawne i etyczne. Problemy te wymuszają na badaczach szukanie nowych, innowacyjnych sposobów na pozyskiwanie i generowanie danych, które mogą być stosowane do trenowania modeli uczenia maszynowego.

Odpowiedzią na te wyzwania jest rosnąca liczba inicjatyw skupionych na tworzeniu otwartych zbiorów danych. Takie inicjatywy promują dzielenie się danymi w społeczności naukowej i technologicznej, co przyczynia się do postępu w dziedzinie sztucznej inteligencji poprzez umożliwienie dostępu do szerokiej gamy zasobów danych.

Automatyzacja procesu oznaczania danych jest kluczowa dla usprawnienia tworzenia zbiorów danych do uczenia nadzorowanego. Stosowanie wcześniejszych iteracji modeli uczenia maszynowego do automatycznego etykietowania obrazów jest jednym ze sposobów na zwiększenie efektywności tego procesu, pozwalając na szybsze przygotowanie dużych zbiorów danych.

Jednakże, automatyczne metody oznaczania wymagają stałej weryfikacji przez ekspertów, aby zapewnić wysoką jakość i dokładność etykiet. Ten proces weryfikacji jest niezbędny, by

unikać propagacji błędów, które mogłyby negatywnie wpłynąć na zdolność modeli do nauki.

Rozwój technologii uczenia maszynowego zależy w dużej mierze od dostępności danych. Postęp w metodach zbierania, oznaczania i stosowania danych ma bezpośredni wpływ na możliwości rozwoju nowych, bardziej zaawansowanych modeli uczenia maszynowego, które mogą uzyskiwać lepszą wydajność przy pracy z problemami, które nas otaczają. Istotne jest podkreślenie potrzeby kontynuacji badań nad metodami poprawy jakości i dostępności danych, jak również nad rozwojem technik automatycznego etykietowania. Tylko poprzez ciągłe doskonalenie tych aspektów, możliwe będzie dalsze przyspieszenie postępu w dziedzinie uczenia maszynowego i przetwarzania obrazu.

W dalszej części rozdziału zostanie przedstawiona specyfika danych, które zostaną zastosowane w ramach badań opisywanych dalej w rozprawie. Opis ten zawierać będzie szczegółowe informacje na temat rodzaju danych, ich źródeł, a także metodyk ich pozyskania. Szczególna uwaga zostanie zwrócona na unikalne cechy zbioru danych, które mają bezpośredni wpływ na procesy uczenia maszynowego, ze szczególnym uwzględnieniem uczenia nadzorowanego. Analiza ta pozwoli na głębsze zrozumienie, jak specyfika stosowanych danych wpływa na efektywność i dokładność modeli uczenia maszynowego stosowanych w przetwarzaniu i klasyfikacji obrazów.

Następnie omówiony zostanie problem dostępu do danych, który stanowi znaczące wyzwanie w dziedzinie uczenia maszynowego i sztucznej inteligencji. Rozdział obejmie przegląd przeszkód, takich jak ograniczenia prawne, etyczne oraz techniczne, które będzie należało pokonać w procesie pozyskiwania danych. Zostanie również szczegółowo opisany cały proces zbierania i oznaczania danych, który zostanie zrealizowany w celu przygotowania zbiorów danych niezbędnych do przeprowadzenia uczenia nadzorowanego. W tym kontekście przedstawione zostaną rozwiązania techniczne i metodyczne zastosowane w pracy, mające na celu maksymalizację efektywności i dokładności procesu oznaczania.

4.1. Proces zbierania danych

W tym podrozdziale przedstawione są etapy pracy nad zbieraniem unikalnych danych do analizy ciosów w walkach bokserskich z zastosowaniem technologii wizyjnych. Znalezienie wartościowych i stabilnych zbiorów danych dotyczących sportów walki, zwłaszcza materiałów nagranych podczas rzeczywistych zawodów i zweryfikowanych przez ekspertów, stanowi znaczące wyzwanie. Dostępne otwarte repozytoria danych często nie oferują zbiorów danych spełniających wymagane kryteria jakości i precyzji, niezbędne do efektywnego trenowania modeli uczenia maszynowego. Dlatego konieczność samodzielnego nagrania danych była nieunikniona, co następnie przełożyło się na szereg działań przygotowawczych i operacyjnych.

Pierwszym krokiem w procesie zbierania danych była selekcja odpowiedniego sprzętu do nagrywania. Do tego celu wybrano 4 sportowe kamery GoPro Hero 8, które ze względu na swoje parametry techniczne i możliwości adaptacyjne, dobrze nadawały się do zadania rejestrowania dynamicznych scen sportowych. Kamery te zapewniają wysoką jakość obrazu, co jest kluczowe dla dokładnego analizowania ruchów zawodników i liczenia ciosów. Dlatego sprzęt zastosowany podczas procesu nagrywania składał się z:

- 4 sportowych kamer GoPro HERO 8, takich jak na rysunku 4.1,

Rysunek 4.1: Model kamery stosowany podczas nagrywania danych

Rysunek 4.2: Model statywu stosowany podczas nagrywania danych

- 4 kart microSD o pojemności 128 GB,
- 4 statywów na kamerę, takich jak na rysunku 4.2,
- 4 banków energii.

Następnie, ważnym etapem było wybranie odpowiedniego wydarzenia sportowego, które umożliwiało nagranie zawodów bokserskich na odpowiednim poziomie rywalizacji w kilku kategoriach wiekowych. Wydarzeniem o takiej charakterystyce była inauguracja bokserskiej ligi młodzików, kadetów oraz juniorów organizowana w 2021 roku w Szczyrku, której plakat znajduje się na rysunku 4.3. Takie wydarzenie pozwoliło na uchwycenie szerokiego spektrum umiejętności i technik stosowanych przez zawodników w różnych kategoriach wiekowych.

Procedura uzyskania zgody organizatora na obecność i możliwość nagrywania na wydarzeniu była kluczowa dla legalności i etyczności całego przedsięwzięcia. Zgoda ta umożliwiła nie tylko sam proces nagrywania, ale także późniejsze stosowanie materiału filmowego do celów badawczych, co stanowiło fundament dla całej rozprawy. Aby tego dokonać niezbędnym było nawiązanie współpracy z Polskim Związkiem Bokserskim, a przede wszystkim z Grzegorzem Proksą (rysunek 4.4 i 4.5) - ówczesnym członkiem zarządu Polskiego Związku Bokserskiego, byłym mistrzem Europy wagi średniej federacji EBU oraz od 2024 roku trenerem reprezentacji Polski w boksie olimpijskim. Rysunek 4.5 zawiera zdjęcie wykonane w dniu nagrywania walk bokserskich i zawiera kolejno (od lewej do prawej) autora niniejszej rozprawy, Pana dr hab. Jana Kozaka prof. UE oraz Pana Grzegorza Proksę.

Po uzyskaniu niezbędnych zgód, kolejnym krokiem było wybranie się na wydarzenie w celu zebrania niezbędnego materiału filmowego. Do tego należało również ustalić pozycje kamer, ostatecznie zdecydowano się na zainstalowanie kamer na statywach za każdym narożnikiem

Rysunek 4.3: Plakat z inauguracji bokserskiej ligi młodzików, kadetów i juniorów, na której odbywał się proces zbierania danych

Rysunek 4.4: Grzegorz Proksa

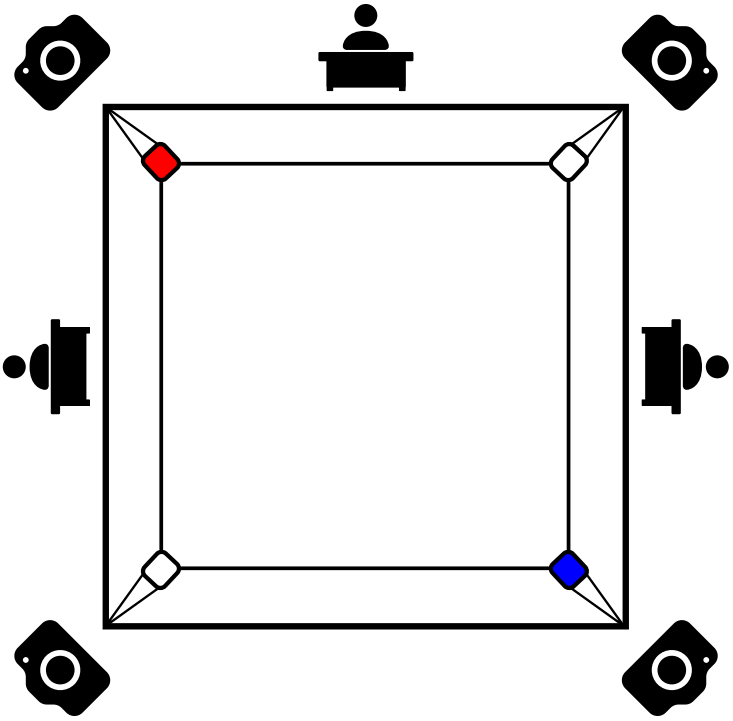
Rysunek 4.5: Zdjęcie z nagrań walk bokserskich

ringu bokserskiego. Takie rozmieszczenie zapewniało kompleksowe pokrycie całego obszaru rywalizacji i umożliwiało nagrywanie zmagania zawodników z różnych perspektyw, co jest niezbędne dla dokładnej analizy i klasyfikacji ciosów. Schemat ringu bokserskiego, pozycji rozmieszczenia sędziów oceniających walkę oraz kamer na statywach został zaprezentowany na rysunku 4.6. Rozważano również umieszczenie kamery nad ringiem bokserskim, lecz sala, na której odbywały się ówczesne zawody nie była do tego przystosowana. Ponadto na podstawie konsultacji z ekspertami dziedzinowymi ustalono, że na innych wydarzeniach sportowych umieszczenie kamery nad ringiem bokserskim również byłoby problematyczne lub też całkowicie niemożliwe.

Po wcześniejszej dyskusji z ekspertami wybrane kamery zostały skonfigurowane tak aby obraz był rejestrowany w rozdzielczości Full HD z szybkością 50 klatek na sekundę. Kamery były umieszczone na statywach na wysokości 1,8m nad parkietem sali gimnastycznej [99, 100]. Proces kalibracji sprzętu oraz środowisko nagrywania walk bokserskich zostały zaprezentowane na rysunku 4.7.

Proces nagrywania wymagał nie tylko precyzji w ustawieniu sprzętu, ale również ciągłej uwagi w celu zapewnienia, że wszystkie ważne momenty zawodów zostaną zarejestrowane. To zadanie okazało się być szczególnie wymagające, biorąc pod uwagę dynamikę walk bokserskich, liczną publiczność oraz konieczność zarządzania wieloma urządzeniami rejestrującymi jednocześnie.

Po zakończeniu zawodów, nastąpił etap archiwizacji zebranego materiału filmowego. Blisko 500 GB danych wymagało starannej organizacji i katalogowania, aby zapewnić łatwy dostęp do poszczególnych fragmentów materiału w dalszej części pracy badawczej. Proces ten był



Rysunek 4.6: Schemat rozmieszczenia kamer oraz sędziów wokół ringu bokserskiego

Rysunek 4.7: Proces kalibracji sprzętu podczas nagrywania

kluczowy dla efektywnego zastosowania danych w analizach i podczas trenowania modeli uczenia maszynowego.

Zebrany materiał filmowy stał się bazą do dalszej analizy i procesu oznaczania, który był konieczny do przeprowadzenia uczenia nadzorowanego. Oznaczanie to polegało na identyfikowaniu i klasyfikowaniu ciosów na nagraniach, co wymagało szczegółowej pracy oraz wiedzy eksperckiej z zakresu walk bokserskich. Proces ten w sposób dokładny został opisany w podrozdziale 4.3.

Proces zbierania danych, choć wymagający i złożony, okazał się być niezbędnym i kluczowym elementem pracy. Zebrane dane nie tylko umożliwiły precyzyjne trenowanie modeli uczenia maszynowego, ale również przyczyniły się do rozwoju wiedzy na temat możliwości zastosowania technologii wizyjnych w analizie sportów walki.

4.2. Wybór i konfiguracja narzędzia do oznaczania danych

Po długiej analizie dostępnych rozwiązań, do procesu oznaczania danych wybrano oprogramowanie CVAT[1] (ang. computer vision annotation tool). Jest to zaawansowane narzędzie do oznaczania obrazów i nagrań wideo o otwartym kodzie źródłowym, rozwijane z myślą o zadaniach związanych z wizją komputerową. To narzędzie jest ukierunkowane na usprawnienie i automatyzację procesu oznaczania danych, co jest kluczowe dla trenowania modeli uczenia maszynowego, zwłaszcza w dziedzinie rozpoznawania obrazów i przetwarzania wideo. Narzędzie to oferuje bogaty zestaw funkcji do oznaczania, umożliwiających dokładne i precyzyjne oznaczanie obiektów w różnorodnych scenariuszach – od prostych zadań klasyfikacyjnych po skomplikowane projekty segmentacji i detekcji obiektów.

Dostępność programu CVAT na darmowej licencji MIT oraz możliwość uruchomienia go na własnej infrastrukturze stanowią jego główne atuty, szczególnie w perspektywie tej rozprawy oraz innych, które wymagają wysokiego poziomu bezpieczeństwa danych. Taka architektura pozwoliła na zachowanie pełnej kontroli nad przetwarzanymi danymi, eliminując ryzyko związane z przesyłaniem wrażliwych informacji na zewnętrzne, nieznane serwery. Ponadto, lokalna instalacja CVAT zapewnia optymalizację czasu wymaganego na transport danych, co jest kluczowe dla efektywności procesu oznaczania, szczególnie w przypadku pracy z dużymi zbiorami danych.

Dzięki dostępności do kodu źródłowego programu CVAT, możliwa była jego konfiguracja na własnej infrastrukturze. Takie podejście do oznaczania zwiększyło bezpieczeństwo przetwarzanych danych oraz szybkość działania narzędzia. Do tego celu zastosowano wirtualną maszynę uruchomioną w serwerowni firmy Google Cloud w Warszawie. Maszyna uzyskała odpowiednie parametry (w tym 4 rdzenie CPU i 15 GB RAM), które zapewniły niezbędną moc obliczeniową oraz minimalizację opóźnień sieciowych. Taka infrastruktura była odpowiednio dopasowana do wymogów procesu oznaczania, gwarantując płynne ładowanie się danych i wysoką szybkość pracy narzędzia, co przyczyniło się do zwiększenia efektywności całego procesu oznaczania.

CVAT został specjalnie zaprojektowany, aby sprostać różnorodnym potrzebom oznaczania

[1] https://github.com/opencv/cvat (Data dostępu: 08.06.2021)

danych wizyjnych, oferując szeroki zakres funkcjonalności, takich jak oznaczanie za pomocą prostokątów, poligonów, linii i punktów. Pozwala to na precyzyjne oznaczanie obiektów o złożonych kształtach i w różnych kontekstach, co jest niezbędne w projektach analizujących ruchy i zachowania w sporcie, takich jak walki bokserskie. Narzędzie to umożliwia również oznaczanie atrybutów obiektów, co pozwala na dodawanie dodatkowych informacji, takich jak typ ciosu lub identyfikacja zawodnika, co jest kluczowe dla głębokiej analizy walk bokserskich.

Podczas procesu oznaczania w rozprawie poświęconej analizie walk bokserskich, CVAT umożliwił szczegółowe oznaczanie każdej klatki wideo, co było niezbędne do precyzyjnego zrozumienia dynamiki walki i identyfikacji kluczowych momentów, takich jak zadanie ciosu. Dzięki intuicyjnemu interfejsowi użytkownika oraz możliwości dostosowania narzędzia do konkretnych potrzeb, proces oznaczania był zarówno efektywny, jak i precyzyjny.

Korzystanie z programu CVAT na własnej infrastrukturze pozwoliło na uniknięcie problemów związanych z opóźnieniami sieciowymi, które mogłyby znacząco wpłynąć na prędkość i płynność pracy nad danymi. Lokalna instalacja narzędzia w serwerowni znajdującej się geograficznie blisko osoby oznaczającej znacząco przyspieszyła czas reakcji narzędzia, co miało bezpośredni wpływ na wydajność całego procesu.

Dodatkowym atutem zastosowania programu CVAT była możliwość pracy zespołowej nad oznaczeniami. Narzędzie to oferuje funkcje zarządzania projektami i zadaniami, co pozwala na efektywne koordynowanie pracy wielu osób, co jest kluczowe w projektach o dużym zakresie, wymagających przetworzenia i oznaczania dużej ilości danych wideo.

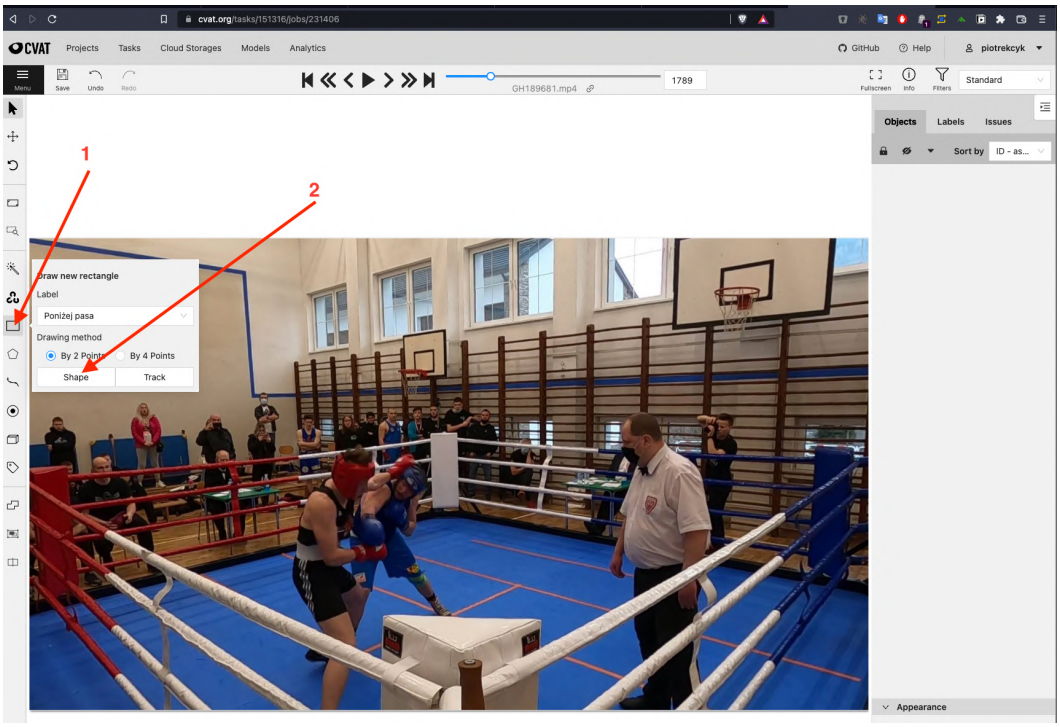
Proces oznaczania w pracy poświęconej analizie walk bokserskich z zastosowaniem programu CVAT wykazał, że odpowiedni wybór i konfiguracja narzędzia do oznaczania danych mogą znacząco wpłynąć na sukces całego przedsięwzięcia badawczego. Zastosowanie otwarto źródłowego narzędzia, dostosowanego do specyficznych potrzeb i wymagań, umożliwiło precyzyjne i efektywne oznaczanie danych, co jest fundamentem dla dalszych analiz i rozwijania modeli algorytmów uczenia maszynowego.

Podsumowując, doświadczenie związane z wyborem i konfiguracją programu CVAT dla potrzeb rozprawy nad analizą walk bokserskich podkreśla znaczenie dostosowania narzędzi do oznaczania danych do konkretnych wymagań badawczych. Zastosowanie tego narzędzia na własnej infrastrukturze, w połączeniu z jego bogatym zestawem funkcji do oznaczeń, zapewniło wysoką efektywność pracy i pozwoliło na zgromadzenie precyzyjnie oznaczonych danych, niezbędnych do realizacji dalszej części rozprawy.

4.3. Proces oznaczania danych

Wynikiem procesu zbierania danych, który został opisany w podrozdziale 4.1 był materiał filmowy zawierający blisko 3 miliony klatek wideo. W celu redukcji materiału przekazanego do oznaczenia przeprowadzono proces selekcji danych. Poprzez mierzenie dystansu między bokserami (proces ten zostanie opisany w podrozdziale 5.2.1) odrzucono te fragmenty wideo, w których zawodnicy pozostają poza zasięgiem ciosów. Redukcja materiału wideo do kluczowych fragmentów zawierających starcia, pozwoliła na znaczące ograniczenie zakresu danych poddawanych procesowi oznaczania. Dzięki temu nie tylko został przyspieszony cały proces oznaczania, ale również zwiększyła się jego efektywność.

Proces oznaczania danych w problemie analizy ciosów w walkach bokserskich za pomocą



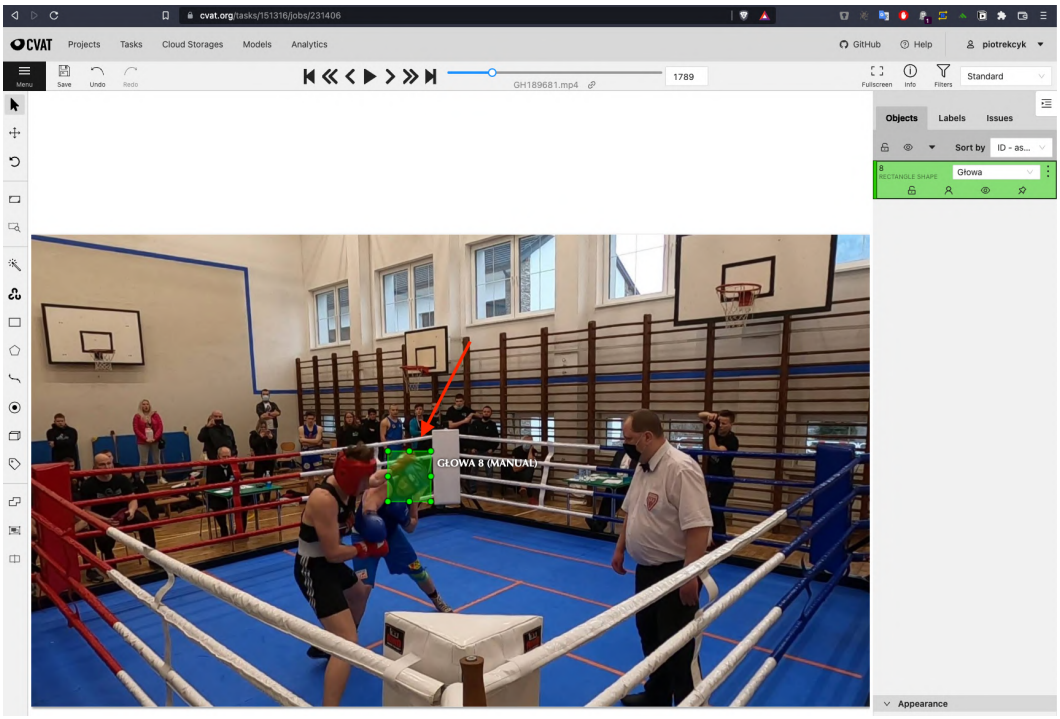
Rysunek 4.8: Prezentacja sposobu oznaczania ciosów: wybranie rodzaju ciosu oraz kształtu którym sędzia będzie oznaczał cios

technologii wizyjnych był kluczowym etapem, który wymagał starannej organizacji oraz posiadania specjalistycznej wiedzy. Po wyborze i instalacji odpowiedniego narzędzia do oznaczania, w tym przypadku CVAT (podrozdział 4.2), konieczne było pozyskanie ekspertów z dziedziny boksu, aby zapewnić wysoką jakość oznaczeń. Licencjonowani sędziowie bokserscy, dzięki swojej wiedzy specjalistycznej, byli idealnymi kandydatami do tego zadania, których ostatecznie udało się pozyskać w celu współpracy.

Zorganizowanie fizycznego szkolenia dla pozyskanych sędziów bokserskich stanowiło kolejny ważny krok. Szkolenie to miało na celu zapoznanie ekspertów z oprogramowaniem oraz specyficznymi wymaganiami technicznymi. W ramach przygotowań stworzono szczegółową instrukcję oznaczania, która służyła jako kompendium wiedzy na temat procesu adnotacji danych.

Instrukcja oznaczania zawierała informacje o sposobie logowania do narzędzia, wyboru zadań oraz szczegółowych instrukcji dotyczących oznaczania poszczególnych ciosów, zarówno tych występujących na pojedynczych klatkach, jak i występujących na przestrzeni kilku klatek. Podkreślono w niej znaczenie przypisywania oznaczeń do odpowiednich typów ciosów, co miało kluczowe znaczenie dla celów badawczych rozprawy.

Oznaczanie ciosów wymagało precyzyjnego nanoszenia kwadratów na miejsca trafień, a w przypadku ciosów trwających kilka klatek, stosowania kwadratów ze śledzeniem. Instrukcja szczegółowo opisywała także nawigację podczas oznaczania, wskazując na sposoby poruszania się między klatkami oraz techniki powiększania obrazu, co miało na celu ułatwienie precyzyjnego oznaczania. Rysunki 4.8 i 4.9 zawierają przykład oznaczania ciosów na nagraniu, który był zawarty w instrukcji przekazanej ekspertom pracującym nad bazą danych.



Rysunek 4.9: Prezentacja sposobu oznaczania ciosów: oznaczenie obszaru, na którym pada cios

Ważnym aspektem instrukcji było zwrócenie uwagi na konieczność zapisywania zmian. Zaproponowano kilka metod zapisu, w tym automatyczny zapis zmian w ustalonych interwałach czasu, co było istotne dla zapewnienia ciągłości pracy i bezpieczeństwa danych. Zaznaczono, że ustawienia zapisywane są w przeglądarce, co oznacza konieczność ponownego ich ustawienia przy zmianie narzędzia pracy.

Rozdzielenie pracy pomiędzy pozyskanych ekspertów było kluczowe dla efektywnej realizacji tego etapu prac. Każdy z sędziów bokserskich otrzymał zestaw nagrań do oznaczenia, co pozwoliło na równomierne rozłożenie obciążenia pracy i maksymalizację wykorzystania specjalistycznej wiedzy każdego z ekspertów.

Proces oznaczania danych przez ekspertów trwał 6 miesięcy, a koordynacja tego procesu wymagała nie tylko ciągłego monitorowania postępów prac, ale również utrzymania motywacji i zaangażowania ekspertów. Regularne spotkania, aktualizacje postępu i wsparcie techniczne były niezbędne, aby zapewnić płynność procesu i rozwiązywać napotykane problemy.

Finalnym etapem procesu oznaczania danych było utrwalenie i zarchiwizowanie całej bazy danych. Ten krok był niezbędny, aby zapewnić trwałość i dostępność zebranych danych dla dalszych etapów badawczych, w tym trenowania modeli algorytmów uczenia maszynowego. Zabezpieczenie danych przed utratą oraz nieautoryzowanym dostępem miało kluczowe znaczenie dla integralności tego i kolejnych etapów prac.

Proces oznaczania był nie tylko technicznym wyzwaniem, ale również wymagał zrozumienia kontekstu bokserskiego, co podkreślało znaczenie wyboru ekspertów z odpowiednią wiedzą dziedzinową. Licencjonowani sędziowie bokserscy, dzięki swoim umiejętnościom i doświadczeniu, byli w stanie zapewnić wysoki poziom precyzji w identyfikacji i klasyfikacji ciosów, co

jest niezbędne dla celów badawczych.

Zaangażowanie ekspertów w proces szkoleniowy, gdzie mieli możliwość zapoznania się z instrukcją oraz praktycznym użyciem narzędzia do oznaczania, pozwoliło na uniknięcie wielu początkowych trudności i przyspieszyło cały proces oznaczania. Dzięki temu, że szkolenie miało formę fizyczną, możliwa była bezpośrednia wymiana doświadczeń i rozwiązywanie wątpliwości na bieżąco.

Proces adnotacji danych był nie tylko czasochłonny, ale również wymagał ciągłej uwagi i precyzji. Dzięki instrukcji oznaczania, sędziowie byli dokładnie poinformowani, jak należy postępować przy każdym typie ciosu, co znacząco przyczyniło się do jednolitości i jakości zgromadzonych danych.

Nawigacja między klatkami oraz możliwości edycji adnotacji w czasie rzeczywistym, wskazane w instrukcji, były kluczowe dla sprawnego oznaczania. Umożliwiły one sędziom szybkie i efektywne przeglądanie nagrań oraz dokładne nanoszenie adnotacji, nawet w przypadku szybkich sekwencji ciosów rozgrywających się na przestrzeni kilku klatek.

Zapisywanie zmian i opcja automatycznego zapisu danych były ważnymi aspektami pracy nad zgromadzonymi danymi. Dzięki tym funkcjom możliwe było minimalizowanie ryzyka utraty danych oraz zapewnienie ciągłości pracy nad oznaczeniami, nawet w przypadku nieoczekiwanych problemów technicznych.

Utrwalenie i zarchiwizowanie danych po zakończeniu procesu adnotacji było równie ważne, co samo oznaczanie. Zapewnienie bezpieczeństwa i dostępności zgromadzonych informacji dla dalszej analizy i badania było priorytetem, który zakończył ten etap prac.

Wnioski płynące z realizacji procesu oznaczania danych podkreślają znaczenie dokładnego planowania, wyboru odpowiednich narzędzi i ekspertów, a także efektywnej organizacji pracy. Każdy z tych elementów miał bezpośredni wpływ na sukces tego etapu prac, umożliwiając zgromadzenie precyzyjnie oznaczonych danych, niezbędnych do dalszych etapów badań nad automatyczną analizą ciosów w walkach bokserskich w niniejszej rozprawie (rozdział 5).

4.4. Opis uzyskanej bazy danych

Utrwalenie zebranej bazy danych po sześciu miesiącach intensywnej pracy nad oznaczeniami nagrań, umożliwiło uzyskanie cennego zbioru do dalszych badań nad automatyczną analizą walk bokserskich. Baza danych zawiera 11 035 (3,53%) przykładów dla klasy „cios” oraz 301 429 (96,47%) dla klasy „nie cios”, co podkreśla niezbilansowany charakter zbioru danych.

Klatki przypisane do klasy „cios” obejmują sytuacje w walce, w których jeden z bokserów uderza swojego przeciwnika. Pozostałe klatki zostały przypisane do klasy „nie cios”, która obejmuje sytuacje w walce, w których bokserzy nie są w bezpośrednim starciu (np. stojąc daleko od siebie) lub są w klinczu. Przykład pojedynczego oznaczenia dla klasy „cios” przedstawiony jest na rysunku 4.10.

Różnice w liczności między klasami są szczególnie istotne w kontekście uczenia maszynowego, gdzie niezbilansowany zbiór danych może wpłynąć na skuteczność i obiektywność modeli klasyfikacyjnych. Ponadto klasa „cios” została dodatkowo podzielona na podklasy w zależności od miejsca uderzenia: 8 140 (73,77%) przypadków dotyczy ciosów w głowę, 990 (8,97%) przypadków ciosów w korpus, a 1 905 (17,26%) przypadków ciosów w blok przeciwnika. Przykłady oznaczeń dla różnych miejsc uderzeń przedstawione są na rysunku 4.11.

Rysunek 4.10: Przykłady oznaczonego ciosu w głowę przez eksperta

Rysunek 4.11: Przykłady różnych typów ciosów oznaczonych przez sędziów bokserskich, gdzie „head” to cios w głowę, „corpus” to cios w korpus, a „block” oznacza cios w blok przeciwnika.

Znacznie większa liczba ciosów w głowę w porównaniu do innych typów ciosów może być odzwierciedleniem naturalnych strategii stosowanych przez bokserów, gdzie dążą oni do zadawania ciosów w obszary potencjalnie gwarantujące szybkie zwycięstwo lub punkty. Z drugiej strony, mniejsza liczba ciosów w korpus może wynikać z większej trudności w efektywnym punktowaniu przez uderzenia w tę część ciała.

Nierównomierność w rozkładzie klas ma kluczowe implikacje dla oceny modeli uczenia maszynowego. Modelom może być trudniej nauczyć się rozpoznawać mniej liczne klasy, co wymaga stosowania specjalnych technik, takich jak ważenie klas, nadpróbkowanie (ang. oversampling) klas mniejszościowych lub podpróbkowanie (ang. undersampling) klas dominujących, aby zapewnić równomierną naukę względem wszystkich przewidywanych kategorii.

Ponadto, wybór miar ocen jakości modeli uczenia maszynowego powinien uwzględniać specyfikę niezbilansowanego zbioru danych. Tradycyjne miary, takie jak dokładność (ang. accuracy), mogą nie odzwierciedlać faktycznej skuteczności modelu w identyfikowaniu rzadszych klas. Dlatego wskazane jest stosowanie miar takich jak precyzja, czułość, miara F1, które lepiej odzwierciedlają zdolność modelu do klasyfikacji (podrozdział 2.6).

Utrwalona i zarchiwizowana baza danych, będąca wynikiem sześciomiesięcznych prac nad identyfikacją i klasyfikacją ciosów w walkach bokserskich, stanowi znaczący wkład w problem analizy sportów walki przy użyciu technologii wizyjnych. Jako fundament dla dalszych badań, zapewnia cenny zasób do trenowania algorytmów uczenia maszynowego w sposób nadzorowany. Ponadto umożliwia ocenę ich skuteczności w oparciu o porównanie z ocenami ekspertów, którzy dokonali oznaczenia pozyskanych wcześniej nagrań. Tym samym, baza ta jest kluczowym elementem w dążeniu do opracowania zaawansowanych narzędzi wspomagających analizę i zrozumienie dynamiki walk bokserskich, otwierając nowe perspektywy dla badań w dziedzinie sportów walki.

Ponadto wystąpiono o interpretację do działu prawnego Uniwersytetu Ekonomicznego w Katowicach w sprawie upublicznienia stworzonej bazy danych w sposób etyczny oraz zgodny z prawem. Zgodnie z interpretacją zadbano o rzetelne przeprowadzenie procesu zamazywania twarzy osób będących na nagraniu w celu ochrony danych osobowych, przy użyciu dostępnych algorytmów. W tym celu wybrano algorytm przedstawiony w pracy [68], która została opublikowana w ostatnich latach na wysoko punktowanej konferencji CVPR. Algorytm został odtworzony na podstawie kodu źródłowego udostępnionego na platformie GitHub[2]. Algorytm został uruchomiony z domyślnymi parametrami, pomijając parametr „score threshold", którego domyślna wartość $0,01$ została nadpisana wartością $0,1$. Po zakończeniu procesu usuwania danych osobowych z zebranych nagrań cała baza wraz z oznaczeniami została opublikowana na platformie Kaggle[3].

[2] https://github.com/damo-cv/mogface (Data dostępu: 22.05.2024)

[3] https://www.kaggle.com/datasets/piotrstefaskiue/olympic-boxing-punch-classification-video-dataset (Data dostępu: 17.06.2024)

5. Prace badawcze

Ten rozdział zawierać będzie opis niezbędnych prac, które zostaną wykonane aby osiągnąć cel główny oraz cele poboczne niniejszej rozprawy. Pierwsze dwa etapy badań nad analizą walk bokserskich będą miały na celu usprawnienie i przyspieszenie procesu oznaczania danych, który będzie kluczowy dla dalszych analiz oraz uczenia algorytmów w sposób nadzorowany. Pierwszy podrozdział 5.1 będzie poświęcony wykrywaniu bokserów na ringu. Zidentyfikowanie zawodników i ich pozycji będzie niezbędne do rozpoczęcia szczegółowej analizy walki. Kontynuując ten kierunek, drugi podrozdział 5.2 będzie poświęcony detekcji starć między bokserami, definiowanych jako momenty, w których zawodnicy znajdują się w bezpośrednim kontakcie lub bliskiej odległości, co sugeruje możliwość wymiany ciosów. Zbudowana wcześniej baza danych (rozdział 4) umożliwia również przejście do kolejnych etapów prac, w których rozpocznie się trenowanie algorytmów uczenia maszynowego w sposób nadzorowany (podrozdział 5.3 i 5.4).

5.1. Wykrywanie bokserów na ringu

Pierwszym krokiem w kierunku wydobywania wiedzy z nagrań walk bokserskich (rozdział 4) jest praca nad podejściem do wykrywania bokserów na ringu. W tym etapie skupiono się na rozwiązaniu problemu wysokiej dynamiki walk bokserskich i obecności licznych elementów w tle, takich jak widownia czy inni zawodnicy. W efekcie udało się zbudować rozwiązanie umożliwiające identyfikację zawodników oraz ich pozycji na ringu, minimalizując jednocześnie wpływ nieistotnych obiektów na dalsze analizy.

W pracy [99] zastosowano kilku etapowy proces do wykrywania bokserów, zaczynając od identyfikacji wszystkich osób na nagraniu, przez wyselekcjonowanie tych znajdujących się na ringu, aż po ostateczne rozróżnienie bokserów od innych osób, takich jak sędziowie czy widownia. Zastosowano zaawansowane techniki przetwarzania obrazu oraz uczenia maszynowego, aby skutecznie filtrować niepożądane elementy i skupić się na zawodnikach. Proces ten wymagał również szczegółowej analizy ubioru zawodników oraz stosowania informacji o kolorach (ściśle określonych w regulaminie boksu olimpijskiego), by dokładnie określić ich położenie na ringu.

Eksperymenty przeprowadzone w ramach tego etapu wykazują wysoką skuteczność proponowanego rozwiązania do wykrywania bokserów, co następnie umożliwia bardziej szczegółowe analizy walk bokserskich bez konieczności manualnego weryfikowania materiału wideo. Dzięki temu możliwe jest automatyczne generowanie statystyk walk. Wyniki eksperymentów, uwzględniające dokładność i pokrycie, potwierdzają zdolność systemu do efektywnego identyfikowania zawodników, nawet w trudnych warunkach dynamicznych walk bokserskich.

Celem tego etapu pracy jest przygotowanie rozwiązania, dzięki któremu możliwe będzie poprawne i szybkie wykrycie zawodników w ringu z pominięciem pozostałych osób rejestrowanych przez kamerę. Pozwoli to na kolejne prace związane z oceną walki bokserskiej.

5.1.1. Wykrywanie bokserów

Trudno uzyskać dostęp do publicznie dostępnych oraz stabilnych nagrań walk bokserskich. Żeby podjąć się problemu wykrywania bokserów na ringu najpierw należało przygotować rzeczywiste nagrania walk bokserskich. Proces ten został opisany w rozdziale 4.

W momencie gdy baza danych nagrań jest już dostępna można przejść do analizy. Przetwarzanie wideo sprowadza się do dzielenia wideo na pojedyncze klatki i analizę każdej z osobna. Okazuje się, że sumarycznie pojedyncze klatki zajmują o wiele więcej miejsca w pamięci, niż wideo w formacie mp4, co jedynie potwierdza skuteczność kompresji danych [114]. Warto o tym pamiętać, ponieważ stworzony potok do przetwarzania danych musi być przygotowany na taki wolumen danych.

Posiadając pojedyncze klatki należy przystąpić do wyszukiwania zawodników na ringu. Jest to problem złożony, ponieważ rejestrując obraz ringu, nagrywane jest również wszystko to, co dzieje się w koło niego. Bardzo często poza ringiem znajduje się widownia oraz inni zawodnicy przygotowujący się do następnej walki. Zatem niezbędne jest ignorowanie wszystkiego co dzieje się poza ringiem, co jest bardzo złożonym procesem.

Wykrywanie zawodników można zacząć od ogólnego wykrywania postaci na obrazie. Ten problem jest w nauce już dobrze znany [9, 111] jednocześnie potwierdzając, że skuteczność gotowych modeli uczenia maszynowego do wykrywania postaci jest na wysokim poziomie. Analizując aktualne podejścia można zauważyć, że w artykule [9] stworzony został nowy przegląd opisujący wydajność kilku popularnych metod głębokiego uczenia (Faster R-CNN [28, 88], R-FCN [18], SSD [67] i YOLOv3 [87]), które zostały zastosowane do wykrywania osób w ruchu ulicznym. Do tego celu zastosowano bazę danych „EuroCity Persons Dataset”[1] zawierającą 283 200 oznaczonych osób na ponad 47 300 zdjęciach zrobionych w 31 miastach w 12 Europejskich krajach. Autorzy uznali, że kluczowym aspektem w wydajności wykrywania osób jest wielkość danych stosowanych w czasie trenowania modelu. Ważna jest również różnorodność, ponieważ na wydajność wykrywania ma wpływ wiele czynników jak między innymi pogoda, która jest zmienna i również uzależniona od regionu na świecie.

Ubiór zawodników na ringu również nie jest w pełni regulowany przez zasady panujące w boksie. Jedynie kask oraz rękawice zawodnika, muszą być podporządkowane do koloru narożnika, który jest kolejno niebieski lub czerwony (podrozdział 3.3.1). Zatem wnioskowanie oparte o kolory należy przeprowadzać uważnie, ponieważ kolor koszulki lub spodenek, które stanowią większość ubioru nie są ściśle określone i ulegają zmianie.

Należy również zauważyć, że na ringu poza bokserami jest również sędzia, który znajduje się bardzo blisko zawodników. Sędzia ma za zadanie rozdzielać bokserów w razie potrzeby, zatrzymywać i wznawiać walkę oraz zwracać uwagę zawodnikom podczas przewinień. Niezbędne jest zatem ignorowanie postaci sędziego.

5.1.2. Metodologia

Warto zauważyć, że przy takim rozwiązaniu, jakie zostało zastosowane w tej rozprawie zawsze któraś z kamer w danym momencie ma dobry kąt widzenia na zawodników. W przypadku gdy zawodnicy zasłaniają się wzajemnie dla dwóch kamer, to dwie pozostałe ustawione są w dobrej pozycji na profile bokserów. Z teoretycznego punktu widzenia cztery kamery

[1] https://eurocity-dataset.tudelft.nl/ (Data dostępu: 03.11.2021)

pozwalają zaobserwować więcej i dokładniej niż trzech oceniających sędziów. Sędziom ponadto podczas oglądania przeszkadzają liny, ponieważ siedzą oni poza ringiem. Rozmieszczenie sędziów zostało naniesione na rysunku 4.6.

Po zebraniu niezbędnego materiału możliwe jest podjęcie się rozwiązania problemu wykrywania zawodników na ringu. Ze względu na złożoność problemu (podrozdział 5.1.1), zaproponowane zostało jego podzielenie na kilka etapów. Poszczególne etapy zostały odzwierciedlone w autorskim algorytmie 5.1, który otrzymuje na wejściu zdjęcie, a jako wynik zwraca listę bokserów oraz ich pozycje. Algorytm w pierwszym kroku wykrywa osoby na zdjęciu, w kroku drugim wybiera osoby będące na ringu, aby w trzecim kroku wybrać jedynie bokserów. Ostatni czwarty krok odpowiedzialny jest za zwrócenie końcowego wyniku.

Algorytm 5.1: Algorytm do wykrywania bokserów na zdjęciu

Input: *image* – jedna klatka z nagrania wideo
Output: *boxers* – lista współrzędnych wykrytych bokserów

```
1 people := find_people_in_image(image);
2 people_on_ring := get_only_people_from_ring(people);
3 boxers := get_only_boxers_from_ring(people_on_ring);
4 result boxers;
```

Podczas pracy nad podejściem wykrywania bokserów na ringu bokserskim napotkano między innymi na problemy:

- odfiltrowania osób, które znajdują się poza ringiem bokserskim;
- odfiltrowania rozgrzewających się poza ringiem zawodników, którzy są już ubrani w rękawice i kaski;
- odfiltrowania sędziego, który jest wszechobecny pomiędzy zawodnikami;
- odfiltrowanie sędziego, który ma ubraną niebieską maseczkę, lub stoi na tle czerwonego lub niebieskiego narożnika;
- wykrywanie bokserów, których sylwetka nie jest w pełni widoczna dla kamery;
- odfiltrowanie osób, które znajdują się blisko ringu skrajnie po lewej lub prawej stronie rejestrowanego obrazu.

Na samym początku do budowy rozwiązania została wybrana próbka danych, którą rozbito na pojedyncze klatki. Następnie zastosowano algorytm do wyszukiwania osób na zadanym obrazie. W kolejnym kroku okazało się to niewystarczające, ponieważ wiele osób stojących za ringiem (trenerzy, widownia czy inni zawodnicy) byli również wykrywani. Niezbędnym zatem było zawężenie obszaru poszukiwań osób, ponieważ wszystko to co dzieje się poza ringiem nie jest warte uwagi w kontekście wykrywania walczących zawodników.

W celu odfiltrowania osób będących poza ringiem zastosowano fakt, że walka toczy się na macie ringu, po której zawodnicy się poruszają. Zatem w celu odrzucenia postaci znajdujących się poza ringiem zastosowano linię odcięcia która została przedstawiona na rysunku 5.2 kolorem czerwonym. Dzięki zastosowaniu linii, akceptowane są jedynie wykryte osoby stojące

na macie ringu bokserskiego. Jak można zauważyć na rysunku 5.2 czerwona linia znajduje się na takiej wysokości, na jakiej kamera rejestruje koniec maty najdalej oddalonego narożnika. Jak można zauważyć przy zastosowaniu linii odcięcia zdarza się akceptować osoby znajdujące się zaraz za ringiem po lewej lub prawej stronie.

Kolejnym problemem jest również fakt, że na macie ringu zawsze są trzy osoby w tym dwóch zawodników. Mowa tutaj o sędzim, który jest w bezpośredniej bliskości zawodników, przez co jego odfiltrowanie jest złożone. Rozwiązanie z zastosowaniem linii odcięcia pozwoliło odrzucić wiele przypadków fałszywie pozytywnych (ang. false positive), lecz nie było wystarczająco dokładne i wymagało usprawnienia.

Do rozwiązania pozostałych problemów zastosowano fakt, że każdy zawodnik walczący w boksie olimpijskim jest ubrany w kask i rękawice bokserskie konkretnego koloru – czerwony lub niebieski. Na tej podstawie zdefiniowano zakres odcieni, nasyceń oraz jasności w modelu HSV (ang. hue, saturation, value) dla koloru niebieskiego oraz czerwonego. Pozwoliło to utworzyć filtr, który pozostawiał na zadanym obrazie jedynie interesujące kolory. Efekt zmiany modelu kolorów z RGB na HSV (podrozdział 1.1) oraz zastosowanie wspomnianego filtra zawiera rysunek 5.1. W ten sposób zostali wykryci zawodnicy, a sędzia wraz z ewentualnymi osobami spoza ringu automatycznie zostali odfiltrowani.

Rysunek 5.1 zawiera wynik procesu wykrywania zawodników, w którym to z zadanego obrazu jest ekstraktowany jedynie kolor czerwony lub niebieski. Na tej podstawie podejmowana jest decyzja czy wykryta osoba jest zawodnikiem, czy też nie.

Rysunek 5.2 zawiera zestawienie zastosowania wszystkich opisanych technik, które finalnie doprowadzają do wykrywania zawodników na ringu. Analizując zdjęcia od lewej do prawej najpierw został przedstawiony czysty widok, na którym nie zastosowano jeszcze żadnej techniki. Następne zdjęcie zawiera już wszystkie wykryte osoby na zdjęciu, następnie na trzecim zdjęciu przedstawiona jest już linia odcięcia. Ostatnie zdjęcie zawiera końcowy wynik, który przedstawia wykryte jedynie dwie osoby będące zawodnikami. Osobom, które znajdują się na zdjęciu rozmazano twarz w celu prywatyzacji.

Rozwiązanie ulepszono również o filtrowanie zawodników, którzy zakończyli już walkę, czyli nie biorących bezpośredniego udziału w pojedynku. W tym celu analizowany jest jedynie kolor znajdujący się w $\frac{1}{8}$ górnej części wykrytej osoby, ponieważ w takiej przestrzeni znajduje się głowa. Takie rozwiązanie pozwala odrzucać zawodników, którzy nie są już ubrani w kaski (zdejmują je od razu po zakończonej walce w swoim narożniku).

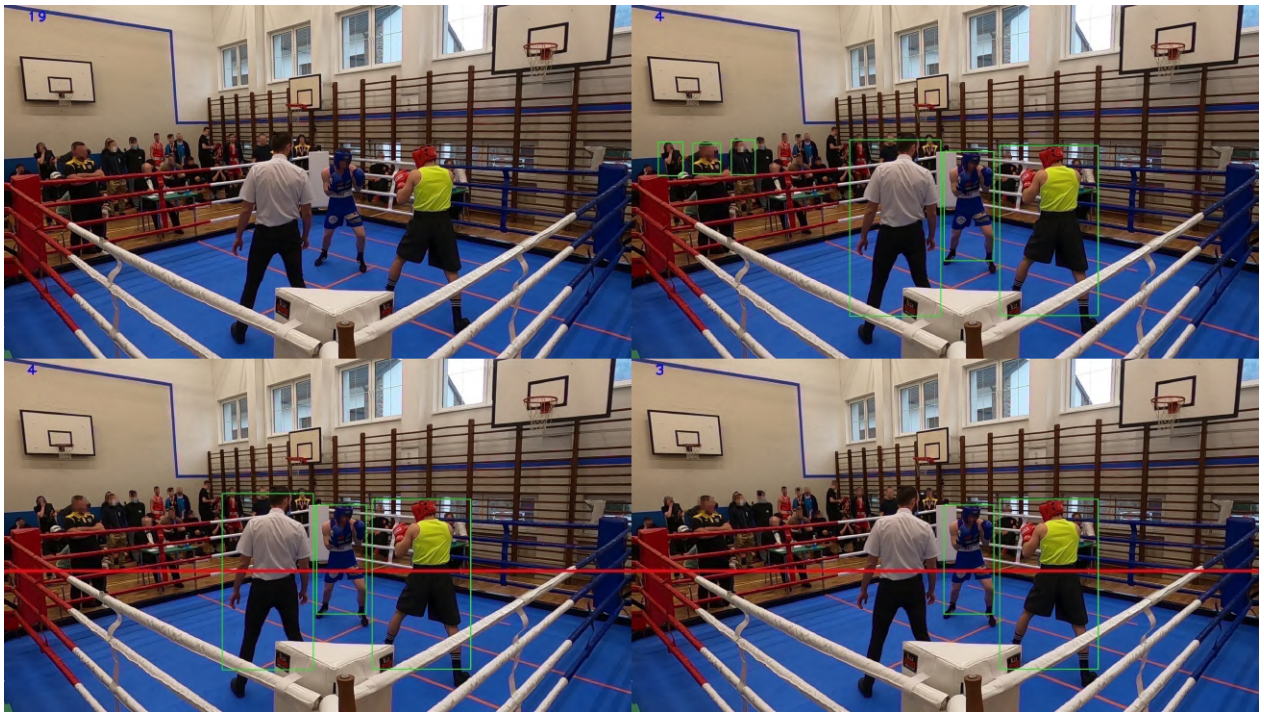
Takim sposobem powyżej opisane rozwiązanie umożliwia wykrywanie zawodników na ringu, znane są dokładne współrzędne położenia bokserów, co umożliwia dokonywanie dalszych analiz i wnioskowania. Podrozdział 5.1.3 zawiera informacje na temat dokładności zaproponowanego rozwiązania.

5.1.3. Eksperymenty

Proponowane rozwiązanie opisane w podrozdziale 5.1.2 dotyczy wykrywania bokserów w ringu bokserskim. W celu przetestowania zaproponowanych rozwiązań przeprowadzone zostały badania eksperymentalne pozwalające na ocenę dokładności i pokrycia każdego z rozwiązań.

Do realizacji tego celu konieczne było przygotowanie rzeczywistego zestawu danych, a następnie przeprowadzenie odpowiednich eksperymentów. Dokładność zosta-

Rysunek 5.1: Wykrywanie koloru niebieskiego i czerwonego w poszukiwaniu zawodników



Rysunek 5.2: Wykrywanie zawodników z zastosowaniem opisanych technik

ła wyliczona na podstawie wzoru (5.1), a pokrycie ze wzoru (5.2), gdzie zmienna: *liczba_poprawnie_wykrytych_bokserow* określa liczbę prawidłowo wykrytych bokserów na ringu; *liczba_wykrytych_osob* określa liczbę wykrytych osób na pojedynczej klatce.

$$dokladnosc\ bokserow = \frac{liczba\_poprawnie\_wykrytych\_bokserow}{liczba\_wykrytych\_osob} \tag{5.1}$$

$$pokrycie\ bokserow = \frac{liczba\_poprawnie\_wykrytych\_bokserow}{2} \cdot dokladnosc\ bokserow \tag{5.2}$$

Nagrania walk bokserskich odbywały się w Polsce na ligowych rozgrywkach śląska dla młodzików, kadetów oraz juniorów. Całe spotkanie trwało cztery godziny i jak już zostało napisane w rozdziale 4 walki były nagrywane czterema kamerami GoPro Hero8. Obraz był rejestrowany w rozdzielczości full HD z częstotliwością 50 klatek na sekundę.

Nagrany materiał został zapisany na komputerze, który posiadał dysk twardy o pojemności 2TB, 8 rdzeniowy procesor o modelu Intel Core i7-2600 CPU @ 3,40GHz oraz 16 GB pamięci RAM. Na tak skonfigurowanym sprzęcie przeprowadzone zostały eksperymenty, które zostały w tym podrozdziale opisane.

Następujące trzy podejścia zostały przetestowane podczas eksperymentów:

Podejście 1 wykrycie wszystkich osób na zdjęciu.

Podejście 2 wykrycie osób stojących na ringu.

Podejście 3 wykrycie bokserów.

Każde z tych podejść ma swoje odzwierciedlenie w krokach przedstawionego algorytmu 5.1, testy zostały przeprowadzone na danych z trzech różnych przedziałów czasowych wynoszących kolejno: 30, 60 oraz 120 sekund.

Wyniki uzyskanych eksperymentów pozwalają na ocenę opisywanych podejść. Dzięki tym badaniom możliwe jest potwierdzenie, że istnieje możliwość na szybkie identyfikowanie zawodników na ringu z jednoczesnym odfiltrowaniem pozostałych osób. Ma to posłużyć do prac związanych z obliczaniem statystyk dotyczących walk bokserskich.

W tabelach 5.1, 5.2 oraz 5.3 zostały zaprezentowane wyniki dla wszystkich trzech podejść dla trzech przedziałów czasowych. Ważne jest, aby mieć na uwadze fakt, że każdy okres czasu obejmuje inną liczbę klatek filmu i tak: 30 sekund, to 1500 obserwacji (50fps x 30s), 60 sekund, to 3000 obserwacji (50fps x 60s), a 120 sekund, to 6000 obserwacji (50fps x 120s). Dlatego przedstawione wyniki są wartościami uśrednionymi.

Jak można zauważyć, w tabeli 5.1 podczas nagrania trwającego 30 sekund, jedna obserwacja (klatka filmu) zawiera średnio 3,90 osoby, z czego 1,89 osób znajduje się na ringu, a 1,12 oznaczone jest, jako bokser. Oznacza to, że już podczas 1500 obserwacji prawie 72% wykrytych osób nie powinna podlegać ocenie podczas walki. Należy też wyjaśnić, że 1,12 wykrytych bokserów wynika z sytuacji, kiedy jeden z zawodników zasłania drugiego lub sędzia zasłania zawodnika. Odniesienie do tego problemu pojawi się jeszcze w dalszych rozważaniach. W przypadku wydłużenia analizy do 60 sekund (3000 obserwacji) sytuacja jest bardzo podobna, ale jeszcze bardziej spada liczba bokserów w ringu – do 0,81, a więc aż

79% osób na jednej obserwacji nie jest bokserami. Najtrudniejsza sytuacja jest w przypadku analizy 120 sekund (6000 obserwacji), ponieważ liczba wszystkich osób wynosi średnio 4,22, a rzeczywistych zawodników 0,43. Przy tym 1,85 osoby jest na ringu – jest to związane przede wszystkim z czasem, kiedy na ringu pojawiają się trenerzy lub następuje wymiana zawodników. Pokazuje to jednak trudność analizy, ponieważ na jednej obserwacji, średnio, niemal 90% osób na obserwacji nie jest bokserem biorącym udział w walce.

Tabela 5.1: Śrenia liczba wykrytych obiektów dla 3 przedziałów czasowych

	Okres czasu		
	30 sekund	**60 sekund**	**120 sekund**
osoby	3,8960	3,8157	4,2168
osoby na ringu	1,8853	1,7217	1,8513
bokserzy	1,1193	0,8080	0,4363

W tabelach 5.2 oraz 5.3 podane są dokładność klasyfikacji i pokrycie w zależności od testowanego podejścia (wyznaczone na podstawie wzoru (5.1) oraz (5.2)). Wyniki te należy analizować łącznie, ponieważ tylko w ten sposób można odzwierciedlić rzeczywistą skuteczność metody.

Na początku można zauważyć, że jednym z podejść można uzyskać bardzo wysoką dokładność wykrycia bokserów (wzór (5.1)). Należy zwrócić jednak uwagę, że podejście 1 w każdym z przypadków uzyskuje bardzo niską dokładność – jest to związane z tym, że wykrywa ono wszystkie osoby podczas każdej obserwacji. Lepsze wyniki uzyskiwane są już w przypadku podejścia 2, ale jedynie w okresie czasu 30 and 60 sekund. W tych przypadkach wykrycie boksera możliwe jest już przy około 50% dokładności. Natomiast podejście 3 daje 100% dokładności, a więc wykrycie zawodnika w przypadku zastosowania tego podejścia pozwala na stwierdzenie, że jest to bokser (dodatkowo biorący udział w walce).

Należy jednak zauważyć, że każda z tych metod ma niskie pokrycie wykrycia bokserów (wzór (5.2)) zaprezentowane w tabeli 5.3. Jest to związane z poruszanym przez nas problemem dotyczącym zasłaniania jednego z zawodników przez sędziego lub drugiego zawodnika. Problem ten należy w przyszłości rozwiązać próbując łączyć nagrania z kilku kamer. Na obecnym etapie jednak należy przedstawić wyniki związane z analizą tylko jednej z kamer. W tym przypadku zdarza się, że wykrywany jest tylko 1 zawodnik, dlatego pokrycie, do wyliczenia którego stosowana jest również dokładność (wzór(5.1) oraz (5.2)) dla podejścia pierwszego spada nawet poniżej 10% (tylko dla 1500 obserwacji udaje się uzyskać 16%). Około dwa razy lepsze wyniki uzyskiwane są w przypadku podejścia 2 (w każdym z zakresów czasu). Natomiast najlepsze wyniki, ponownie, uzyskane zostały dla podejścia 3. Zastosowanie wszystkich rozwiązań przedstawionych w tym podrozdziale pozwala na uzyskanie pokrycia na poziomie 56% dla 1500 obserwacji, 40% dla 3000 obserwacji i 22% dla 6000 obserwacji (120 sekund). Biorąc pod uwagę opisane problemy jest to wynik satysfakcjonujący i pozwalający na efektywną pracę z tak przygotowanym materiałem badawczym.

5.1.4. Podsumowanie

Celem prac opisanych w tym podrozdziale było zaproponowanie rozwiązania, które pozwoli wykrywać zawodników na ringu bokserskim. Aby rozpocząć prace nad rozwiązaniem niezbędne

Tabela 5.2: Dokładność wykrycia bokserów w analizowanych podejściach

	Okres czasu		
	30 sekund	**60 sekund**	**120 sekund**
Podejście 1	0,2880	0,2118	0,1035
Podejście 2	0,5937	0,4693	0,2357
Podejście 3	1,0000	1,0000	1,0000

Tabela 5.3: Pokrycie wykrycia bokserów w analizowanych podejściach

	Okres czasu		
	30 sekund	**60 sekund**	**120 sekund**
Podejście 1	0,1612	0,0856	0,0226
Podejście 2	0,3323	0,1896	0,0514
Podejście 3	0,5597	0,4040	0,2182

było zgromadzenie odpowiedniego materiału. Aby tego dokonać nagrano rzeczywiste walki bokserskie, a następnie przygotowano cały zestaw danych.

Problem wykrywania zawodników na ringu jest złożony i został podzielony na trzy etapy. Najpierw zostały wykryte wszystkie osoby na zadanym obrazie, następnie metodą eliminacji zostały odfiltrowane osoby poza ringiem, w kolejnym kroku odfiltrowano sędziego. W taki sposób zaproponowane rozwiązanie dostarcza informacji na temat liczby zawodników na ringu oraz współrzędnych ich położenia.

Przeprowadzone doświadczenia potwierdzają, że zaproponowane rozwiązania pozwalają na zdecydowaną poprawę dokładności wykrywania bokserów na ringu. Od prostego podejścia, które wykrywało wszystkie osoby na klatce filmu udało się uzyskać efekt, w którym zostali wykryci jedynie bokserzy na ringu. Pewnym kłopotem pozostaje jeszcze pokrycie wykrycia bokserów, ponieważ nie zawsze wykrywani są wszyscy zawodnicy. Dzieje się tak w sytuacji, kiedy zawodnik jest zasłonięty przez drugiego zawodnika lub sędziego. Jest to największa wada proponowanego podejścia.

Wyniki uzyskane w tym etapie prac stanowią fundament dla dalszych analiz, które zostały przedstawione w kolejnych rozdziałach. W szczególności wykrywanie bokserów jest podstawą dla obliczania odległości pomiędzy zawodnikami, a ten problem został opisywany w podrozdziale 5.2.

5.2. Pomiar dystansu pomiędzy bokserami i wykrywanie starć

Etap ten skupia się na stosowaniu wizji komputerowej do analizy walk bokserskich, konkretnie na wykrywaniu momentów bezpośredniego kontaktu między bokserami, co jest kluczowym elementem w ocenie aktywności zawodników podczas walki. Stosując technologię rozpoznawania obrazu, została stworzona metoda pozwalająca na identyfikację i analizę starć bokserów, definiowanych jako momenty gdy zawodnicy są na tyle blisko siebie, że możliwe jest zadanie ciosów. Celem tej metody jest nie tylko dostarczenie bardziej szczegółowych

statystyk, ale również znaczące usprawnienie procesu analizy walk poprzez automatyczne odfiltrowywanie nieistotnych fragmentów walki, gdzie zawodnicy znajdują się poza zasięgiem ciosów.

Podejście to opiera się na wcześniejszym etapie prac dotyczących wykrywania obecności bokserów na ringu 5.1. Przy użyciu zbioru danych z rzeczywistych walk bokserskich (rozdział 4), zrealizowano pomiar odległości euklidesowej między zawodnikami jako podstawę do wykrywania starć. Proces ten pozwala na wyeliminowanie fragmentów walki, w których zawodnicy są zbyt daleko od siebie, aby mogło dojść do wymiany ciosów, co skutecznie redukuje objętość danych potrzebnych do analizy i znacząco przyspiesza proces oznaczania walk.

Eksperymenty przeprowadzone na nagraniach z prawdziwych zawodów bokserskich wykazują, że w ponad 70% czasu trwania walki bokserzy nie angażują się w bezpośredni kontakt, co pozwala na znaczącą redukcję materiału wideo podlegającego analizie. Przez zastosowanie opracowanej metodyki, możliwe jest skupienie się jedynie na tych fragmentach nagrania, gdzie dochodzi do bezpośredniej konfrontacji między zawodnikami, co jest szczególnie wartościowe w kontekście dalszych analiz takich jak klasyfikacja i ocena ciosów.

5.2.1. Wykrywanie starć

Zawody bokserskie charakteryzują się dużą liczbą przerw pomiędzy walkami i rundami. Podczas oglądania materiału z całych zawodów wydaje się, że przez długi czas na ringu bokserskim nic się nie dzieje. Wykrywanie starć jest etapem wstępnego przetwarzania przed wykryciem i klasyfikacją ciosów na nagraniu. Z obserwacji wynika, że przez około 70% czasu (w nagraniach z całych zawodów) bokserzy nie są zaangażowani w walkę wręcz.

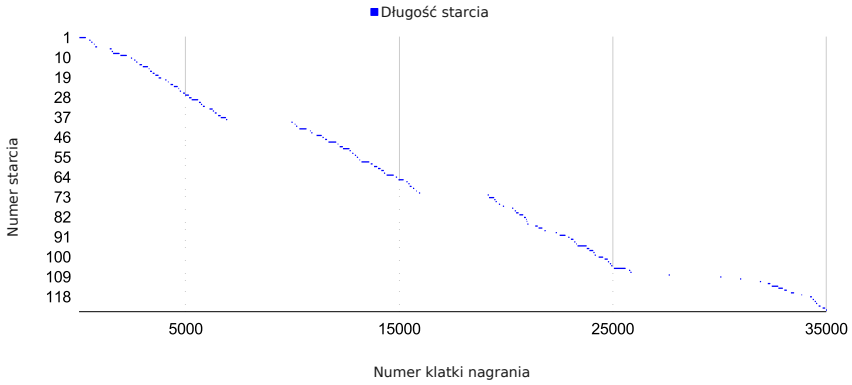
Boks to bardzo dynamiczny sport, który wymaga kamer o wysokiej rozdzielczości i dużej liczbie klatek na sekundę, aby uchwycić szczegóły. Korzystanie z tak wysokiej jakości podczas rejestrowania obrazu przekłada się na dużą złożoność obliczeniową. Zatem podejście do wykrywania starć, które zmniejsza ilość przetwarzanych danych, jest niezwykle istotne.

Metodologia wykrywania starć oparta jest na wcześniejszym podejściu do wykrywania bokserów (podrozdział 5.1 oraz [99]). Wykrywanie bokserów bazuje na technice wykrywania osób [9, 111] oraz podejściu opartym na kolorach do filtrowania sędziego i osób poza ringiem bokserskim.

Miara euklidesowa z równania (1.6) jest stosowana do obliczania odległości między bokserami na ringu. Algorytm wykrywa starcie, gdy odległość między wykrytymi bokserami jest niewielka. Starcie to potencjalna sytuacja, w której może dojść do ciosu między bokserami.

Ponadto do eksperymentów wybrano 12-minutowy materiał zawierający 35 000 klatek. Materiał obejmuje 3 rundy jednej walki i jedną rundę następnej. Pomiędzy rundami 3 i 4 wykryto kilka starć, co można zauważyć na rysunku 5.3. Pomiędzy walkami zawodnicy witają się, dziękują sobie i odbierają nagrody, dlatego w takich sytuacjach mogą pojawiać się fałszywie pozytywne przypadki wykrycia starć.

Rysunek 5.3 zawiera wykryte starcia w przetwarzanym nagraniu wideo oraz ich długość trwania. W wybranym materiale wykryto starcia na 14 600 klatkach, co oznacza, że przez 41% nagrania bokserzy znajdują się w sytuacji walki wręcz. W ten sposób można odfiltrować pozostałe 59% nagrań. Luki na wykresie reprezentują przerwy między rundami i sytuacje, w których bokserzy nie znajdują się w bezpośredniej bliskości.



Rysunek 5.3: Wykres przedstawiający wykryte starcia oraz ich długość

5.2.2. Podsumowanie

Celem tego etapu prac było zaproponowanie podejścia do wykrywania starć w nagraniach walk bokserskich. W wyniku eksperymentów potwierdzono, że metodologia wykrywania starć może być stosowana w części wstępnego przetwarzania nagranego materiału, co pozwala skutecznie zmniejszyć dużą ilość danych jednocześnie nie tracąc istotnych informacji.

Przedstawiony etap prac stanowi wkład w rozwój metod analizy w sportach walki, otwierając nowe perspektywy dla badań nad boksem oraz potencjalnie innymi dyscyplinami sportowymi, gdzie kluczowe jest rozpoznawanie momentów bezpośredniego kontaktu między zawodnikami.

Opracowana metodologia pozwoliła zredukować ilość przekazanych danych do oznaczania, co znacznie przyczyniło się do przyspieszenia tego procesu (podrozdział 4.3). Dalsze etapy prac w niniejszej rozprawie bazują na stworzonej bazie danych jednocześnie stosując nadzorowane podejścia do trenowania algorytmów uczenia maszynowego.

5.3. Klasyfikacja ciosów

Kolejny etap badań poświęcony jest zastosowaniu konwolucyjnych sieci neuronowych (ang. convolution neural networks, CNN) do klasyfikacji ciosów w boksie olimpijskim, bazując na danych wizyjnych z kamer RGB. Celem jest opracowanie metody pozwalającej na automatyczne klasyfikowanie ciosów zadawanych przez zawodników, z zastosowaniem już oznaczonej bazy danych (rozdział 4). Licencjonowani sędziowie dokonali klasyfikacji każdego ciosu, przypisując go do jednej z trzech kategorii: cios w głowę, cios w korpus lub cios w blok. Zastosowanie tak szczegółowo oznaczonej bazy danych ma na celu stworzenie skutecznego modelu klasyfikującego, zdolnego do dokładnego rozróżniania poszczególnych typów ciosów.

W badaniu tym, sieci CNN są stosowane do opracowania modelu zdolnego do klasyfikowania ciosów na podstawie obrazu z kamer RGB umieszczonych naprzeciwko ringu bokserskiego.

W celu poprawy skuteczności modelu stosuje się technikę augmentacji danych, która polega na sztucznym zwiększeniu różnorodności zestawu danych poprzez wprowadzenie losowych, ale realistycznych modyfikacji do przetwarzanych obrazów. Proces ten ma na celu zapobieganie

nadmiernemu dopasowaniu modelu do danych treningowych oraz zwiększenie jego zdolności generalizacji na nowych, niewidzianych wcześniej danych.

Eksperymenty przeprowadzone na zestawie danych zawierającym przykłady ciosów wykazują, że proponowany model CNN osiąga wysoką skuteczność w ich klasyfikacji. Dla trzech klas ciosów (głowa, korpus, blok) uzyskuje się kolejno 94%, 84% i 81% wyniku miary F1, co potwierdza wysoką precyzję i pokrycie klasyfikacji. Wyniki te stanowią solidną podstawę do dalszego rozwoju systemów analizy wideo w boksie.

Analiza wpływu augmentacji danych na wydajność modelu wykazuje znaczącą poprawę w klasyfikacji ciosów, szczególnie w przypadku klas trudniejszych do rozróżnienia. Wzrost wyników miary F1 dla wszystkich kategorii ciosów podkreśla efektywność augmentacji danych jako techniki zwiększającej skuteczność modelu CNN w zadaniach klasyfikacji obrazów.

5.3.1. Inne podejścia

Wykrywanie i śledzenie obiektów to jedne z najważniejszych zagadnień w wizji komputerowej [90, 120, 125]. W przemyśle systemy wizyjne automatycznie monitorują zachowanie pracowników i środków ochrony osobistej w celu poprawy zdrowia i bezpieczeństwa podczas pracy [65, 90, 131]. Istnieją również rozwiązania monitorujące jakość produkowanych elementów [6, 30]. Przetwarzanie obrazu jest szeroko stosowane w ruchu drogowym [14, 41], obszarach publicznych (takich jak stacje kolejowe) [14, 16], przemyśle budowlanym [75, 90] lub sporcie [107]. Śledzenie obiektów zapewnia szerokie możliwości badania zachowania wykrytych obiektów w wielu różnych obszarach naszego życia, takich jak zapobieganie upadkom w szpitalach [74], wykrywanie napaści w przestrzeni publicznej [14] lub analizowanie ruchu graczy na boisku [108].

W wielu dyscyplinach sportowych kamery są stosowane przez zaawansowane komputerowe systemy wizyjne, które mogą dostarczyć dodatkowych cennych informacji o nagrywanej scenie. Dlatego też na stadionach sportowych zainstalowano dziesiątki kamer, które automatycznie śledzą sportowców i analizują ich ruch w celu dalszej analizy [108]. Takie dane mają szeroki zakres interesariuszy, takich jak sędziowie, komentatorzy sportowi lub publiczność. Również trenerzy powszechnie stosują te dane do analizy wyników całej drużyny lub poszczególnych sportowców [20, 64, 133].

W sportach walki pojawiły się dwa podejścia do analizy wyników bokserów, które można podzielić ze względu na rodzaj urządzeń używanych do analizy bokserów. Jedno z nich stosuje urządzenia do ubrania, takie jak czujniki tekstylne [124], czujniki wbudowane w sprzęt bokserów (rękawice bokserskie i ochraniacze głowy) [31, 118] lub czujniki na plecach bokserów [118]. Drugie podejście nie koliduje ze sprzętem bokserskim przy użyciu kamery RGB-D [7, 45, 46] lub statycznej kamery RGB naprzeciwko ringu bokserskiego [116], tym samym nie narażając zdrowia boksera.

Podczas pracy z interesariuszami pomysł zastosowania urządzeń lub wszelkich czujników instalowanych na ciele zawodnika został odrzucony, głównie z powodu panujących restrykcji w przepisach bokserskich [46] (rozdział 3). Aby na odległość zmierzyć wydajność oraz badać zachowanie bokserów można zastosować kamerę. Kamera nad ringiem (używana przez [7, 45, 46]) rejestruje czystsze dane bez publiczności i zaciemnień między bokserami niż kamera naprzeciwko ringu bokserskiego. Jednak nie wszystkie ringi bokserskie są przygotowane do instalacji urządzeń nad nimi, na ringu, na którym dane były zbierane (rozdział 4) również nie

było takiej możliwości. Dlatego zdecydowano się użyć statycznych kamer RGB, takich jak w pracy [116], gdzie autorzy stworzyli system śledzenia rękawic bokserów.

Uczenie maszynowe jest powszechnie stosowane do rozwiązywania problemów związanych z klasyfikacją. Istnieje wiele algorytmów, takich jak drzewa decyzyjne (ang. Random Forest) lub maszyny wektorów nośnych (ang. Support Vector Machines, SVM) używane przez autorów [46] do klasyfikacji ciosów na podstawie danych o głębi. Ponadto w pracy [12] zastosowano sieci CNN (ang. Convolutional Neural Network) do klasyfikacji działań w sporcie na podstawie obrazów. CNN to jedna z najpopularniejszych struktur głębokich sieci neuronowych, która składa się z wielu warstw, w tym warstwy konwolucyjnej, warstwy nieliniowej, warstwy łączącej i warstwy w pełni połączonej. CNN jest z powodzeniem stosowana w problemach uczenia maszynowego, takich jak klasyfikacja obrazów lub przetwarzanie języka naturalnego (ang. Natural Language Processing, NLP) [1]. Dlatego zdecydowano się zastosować sieć CNN do klasyfikacji ciosów z obrazów.

5.3.2. Metodologia

Głównym celem tego etapu pracy jest zaproponowanie rozwiązania do klasyfikacji ciosów w boksie olimpijskim. Zdecydowano się użyć kamer RGB rozstawionych wokół ringu bokserskiego, aby uchwycić scenę. Ze względu na brak takich danych, konieczne było przygotowanie własnego zestawu uczącego. Rozdział 4 zawiera cały proces budowania bazy danych, która została zastosowana do dalszych eksperymentów. Do klasyfikacji zastosowano model CNN, do którego dodatkowo dodano etap augumentacji danych w celu zróżnicowania zbioru danych i zapobieganiu nadmiernemu dopasowaniu [35, 130].

W celu klasyfikacji uderzeń zdecydowano się użyć konwolucyjnej sieci neuronowej (CNN), która jest popularną strukturą głębokich sieci neuronowych, powszechnie stosowaną w problemach klasyfikacji obrazów. Wytrenowana sieć neuronowa została zaprojektowana do wykrywania trzech klas ciosów:

1. Head - cios w głowę przeciwnika.
2. Corpus - cios w korpus przeciwnika.
3. Block - cios w rękawice przeciwnika.

Wydajność klasyfikacji została zaprezentowana w podrozdziale 5.3.3 i oceniana za pomocą następujących wskaźników: dokładność (2.19), precyzja (2.20), pokrycie (2.21) i miara F1 (2.23) [29, 95].

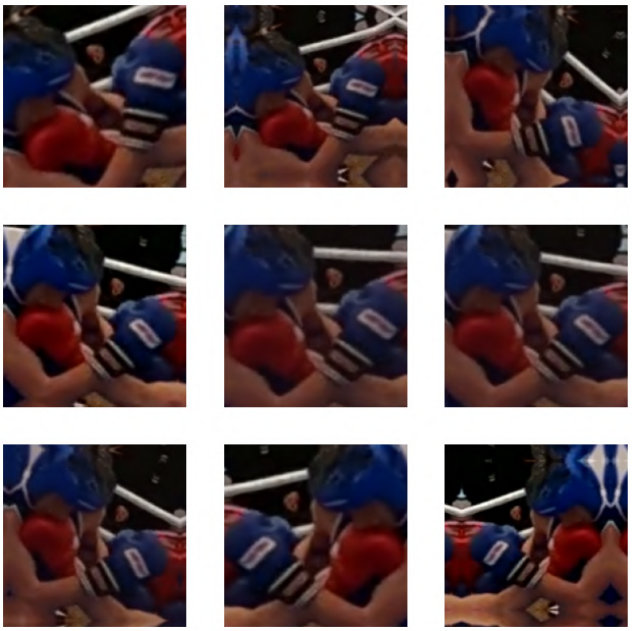
Proces etykietowania danych dostarczył 5115 obrazów z ciosami podzielonymi na trzy klasy; 3575 (około 70%) przypisano do klasy „Head”; 458 (około 9%) przypisano do klasy „Corpus”, a 1082 (około 21%) przypisano do klasy „Block”. Aby zdywersyfikować zbiór danych i zapobiec nadmiernemu dopasowaniu [35, 130], zastosowano proces rozszerzania danych. Technika ta stosuje losowe, ale realistyczne przekształcenia na każdym przetwarzanym obrazie:

1. losowe odbicie lustrzane (ang. random flip) - operacja losowego przerzucania obrazów co tworzy efekt lustrzanego odbicia.
2. Losowy obrót (ang. random rotate) - operacja losowego obracania obrazów.

3. Losowe przybliżenie/oddalenie (ang. random zoom) - operacja losowego powiększania obrazów.

4. Losowy kontrast (ang. random contrast) - operacja losowej modyfikacji kontrastu obrazów.

Rysunek 5.4 zawiera kilka przekształceń obrazu z procesu rozszerzania danych.



Rysunek 5.4: Przekształcenia rozszerzające dane

5.3.3. Eksperymenty

Wyniki eksperymentów zostały uzyskane na komputerze z 16-rdzeniowym procesorem Intel Core i9-11900K @ 3,50 GHz, 64 GB pamięci RAM i kartą graficzną Nvidia Geforce GTX 1080Ti, na systemie operacyjnym Ubuntu.

Podczas eksperymentów zostały wytrenowane dwa klasyfikatory oparte na sieciach neuronowych i architekturze CNN. Pierwszy opierał się na zestawie danych bezpośrednio z procesu etykietowania, bez żadnych przekształceń. Drugi opierał się na tym samym zestawie danych, ale dodatkowo zawierał etap wstępnego przetwarzania z rozszerzeniem danych. Krok z transformacjami obrazu został zaproponowany w celu zróżnicowania zbioru danych i uniknięcia problemów z nadmiernym dopasowaniem, szczegóły transformacji zostały opisane w podrozdziale 5.3.2. Obie sieci na warstwie wejściowej otrzymały kolorowe obrazy w rozdzielczości 180x180 px, dodatkowo klasyfikatory zostały wytrenowane z zastosowaniem optymalizatora Adam, a entropia krzyżowa została zastosowana jako funkcja straty.

Klasyfikatory zostały ocenione za pomocą czterech wskaźników, które zostały opisane w podrozdziale 5.3.2. Tabele 5.4 i 5.5 zawierają wartości wskaźników precyzji, pokrycia i miary F1 dla każdej klasy. Z kolei rysunek 5.5 i rysunek 5.6 zawierają proces uczenia oraz walidacji modelu prezentując dokładność oraz wartość funkcji straty.

Pierwszy klasyfikator uzyskał 82% dokładności i osiągnął najlepsze wyniki precyzji (85%), pokrycia (92%) i miary F1 (88%) dla klasy „Head". Mimo że klasa „Block" zawierała więcej przykładów niż klasa „Corpus", była najtrudniejsza do sklasyfikowania, jej precyzja wyniosła 69%, a pokrycie 55%. Klasa „Corpus" ma o ponad połowę mniej przykładów niż klasa „Block" i uzyskuje 81% precyzji i 68% pokrycia.

Drugi klasyfikator został wytrenowany na rozszerzonym zestawie danych i uzyskał dokładność 90%. Podobnie jak w przypadku pierwszego klasyfikatora, klasa „Head" osiągnęła najlepsze wyniki precyzji (93%), pokrycia (94%) i miary F1 (94%). Klasyfikacja klasy „Block" była nadal najtrudniejsza, biorąc pod uwagę wynik miary F1, który był najniższy i wynosił 81%. Niemniej jednak dysproporcja w wyniku F1 między klasami „Corpus" i „Block" została zmniejszona o 10 punktów procentowych. Znaczący wzrost wyniku F1 dla klasy „Block" był spowodowany wzrostem precyzji o 12 punktów procentowych i pokrycia o 26 punktów procentowych.

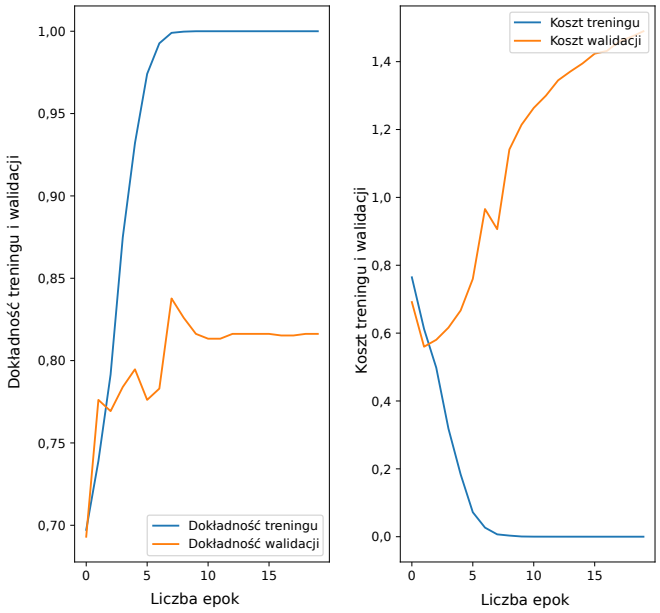
Proces rozszerzania danych miał znaczący wpływ na wydajność klasyfikacji. Biorąc pod uwagę wynik miary F1, został on zwiększony o 6 punktów procentowych dla klasy „Head", o 10 punktów procentowych dla klasy „Corpus" i o 20 punktów procentowych dla klasy „Block". Miało to również pozytywny wpływ na proces uczenia się, rysunek 5.5 zawiera wykresy procesu uczenia się bez procesu rozszerzania danych, a rysunek 5.6 zawiera wykresy po tym procesie. Porównując te wykresy można stwierdzić, że proces rozszerzenia danych skutecznie zmniejszył problem nadmiernego dopasowania oraz zwiększył stabilność klasyfikatora. Można to zaobserwować poprzez porównanie przestrzeni pomiędzy linią określającą dokładność treningu, a linią określającą dokładność walidacji, im ta przestrzeń jest mniejsza, tym mniejszy występuje problem z przetrenowaniem, a sam model cechuje się lepszą zdolnością do generalizacji. Takie samo porównanie ma zastosowanie na wykresie zawierającym koszt treningu oraz walidacji, im mniejsza przestrzeń pomiędzy krzywymi, tym mniejszy problem z nadmiernym dopasowaniem. Jak można zauważyć proces rozszerzania danych znacznie zniwelował problem z przetrenowaniem modelu poprawiając proces uczenia (krzywe na rysunku 5.6 są o wiele bliżej siebie w porównaniu z rysunkiem 5.5) jak i same wskaźniki wydajności (tabela 5.5).

Tabela 5.4: Jakość klasyfikacji

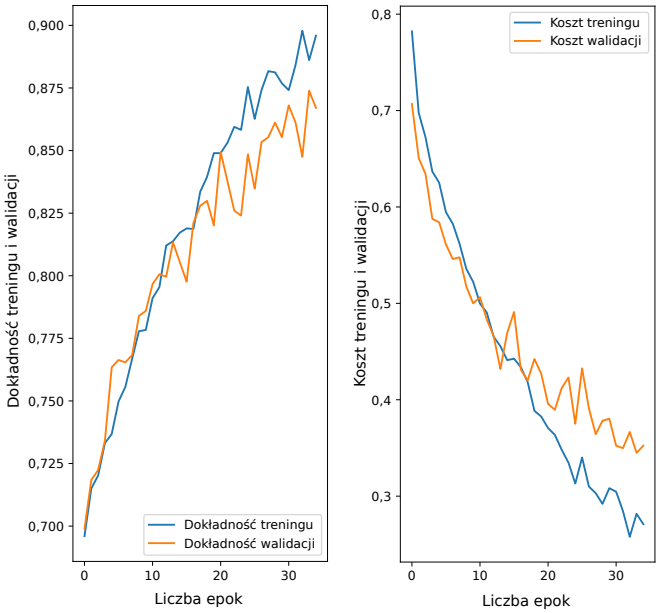
Klasa	Precyzja	Pokrycie	F1
Head	**0,85**	**0,92**	**0,88**
Corpus	0,81	0,68	0,74
Block	0,69	0,55	0,61

Tabela 5.5: Jakość klasyfikacji z procesem rozszerzania danych

Klasa	Precyzja	Pokrycie	F1
Head	**0,93**	**0,94**	**0,94**
Corpus	0,90	0,78	0,84
Block	0,81	0,81(↑**26%**)	0,81(↑**20%**)



Rysunek 5.5: Proces uczenia



Rysunek 5.6: Proces uczenia z procesem rozszerzania danych

5.3.4. Podsumowanie

Uzyskane wyniki eksperymentów dowiodły, że proponowane rozwiązanie do klasyfikacji ciosów przy użyciu kamer RGB jest możliwe i osiąga wysoką wydajność. W eksperymentach dwa klasyfikatory bazujące na sieci neuronowej CNN zostały przeszkolone i przetestowane. Zmierzono również wpływ procesu rozszerzania danych na wydajność klasyfikacji. Klasyfikatory przewidywały trzy klasy ciosów, a najlepszy z nich uzyskał dla nich kolejno 94%, 84% i 81% wyniku miary F1.

Podsumowując, ten etap prac (również opublikowany [98]) dostarczył istotnego wkładu w rozwój automatycznej analizy walk bokserskich, prezentując skuteczną metodę klasyfikacji ciosów z zastosowaniem konwolucyjnych sieci neuronowych (CNN) na podstawie danych wizyjnych. Stosowanie statycznych kamer RGB do zbierania materiału wideo oraz zastosowanie nieinwazyjnego podejścia, które nie wymaga modyfikacji ubioru zawodników, stanowi ważny krok w kierunku przystosowania technologii do obowiązujących przepisów boksu olimpijskiego. Badanie to otwiera drogę do dalszych prac nad automatyczną analizą walk bokserskich, sugerując możliwość rozszerzenia na inne dyscypliny oraz rozwój bardziej złożonych systemów klasyfikacyjnych, które mogą w przyszłości służyć do jeszcze dokładniejszej analizy technik i strategii stosowanych przez zawodników.

Na podstawie przeprowadzonych badań oraz uzyskanych wyników, należy podkreślić, iż skuteczne stosowanie opracowanego klasyfikatora ciosów bokserskich, wymaga zastosowania dodatkowego klasyfikatora do identyfikacji momentów na nagraniu, w których zostaje zadany cios. Dopiero takie połączenie dostarczy zawodnikom oraz ich trenerom wartościowych informacji o przebiegu walki. Z tego powodu kolejne etapy pracy w ramach niniejszej rozprawy skoncentrowane zostały na rozwoju wspomnianego klasyfikatora.

5.4. Wykrywanie ciosów

W tej części badań skoncentrowano się na wykrywaniu ciosów w boksie olimpijskim z zastosowaniem statycznej kamery RGB, co stanowi istotny wkład w problematykę analizy sportów walki za pomocą wizji komputerowej. Celem tej części jest automatyzacja procesu wykrywania momentów, w których pada cios, co ma znaczące zastosowanie w analizie wydajności zawodników.

Podejście do badania opiera się na klasyfikacji pojedynczych klatek wideo jako zawierających cios (klasa „punch") lub nie zawierających ciosu (klasa „not punch"), z zastosowaniem sieci CNN. Z przeprowadzonych badań wynika, że kluczowym wyzwaniem jest efektywna klasyfikacja pomimo faktu, że obiekt determinujący klasę decyzyjną zajmuje bardzo małą część rejestrowanego obrazu. W tym przypadku obszar zadawania ciosu zawiera się na poniżej 1,5% powierzchni całej klatki. Z tego powodu oceniane są różne metody wstępnego przetwarzania klatek wideo, mające na celu poprawę skuteczności klasyfikatora do wykrywania ciosów. Testowane są trzy podejścia do segmentacji klatek wideo przed przekazaniem ich do klasyfikatora: ekstrakcja kolorów, odejmowanie statycznego tła oraz metoda hybrydową łączącą oba podejścia. Do metody bazowej stosuje się oryginalne klatki wideo, bez żadnego przetwarzania wstępnego.

Eksperymentalne potwierdzenie skuteczności proponowanych metod wstępnego przetwarzania wykazuje, że największą efektywność w poprawie klasyfikacji klatek zapewnia podejście

oparte na odejmowaniu statycznego tła. Podejście to umożliwia osiągnięcie 95% zbalansowanej dokładności, co podkreśla jego zdolność zastosowania w dalszych etapach rozprawy.

Na podstawie przeprowadzonych badań zaprezentowany jest działający system analizy scen bokserskich, zdolny do oznaczania zawodników, etykietowania klatek z wykrytymi starciami i ciosami oraz zliczania sytuacji zawierających ciosy. System ten ma zdolność do automatycznego generowania skrótów walk oraz ułatwienia pracy trenerów i komentatorów sportowych poprzez dostarczanie szczegółowych statystyk dotyczących przebiegu pojedynków.

5.4.1. Inne podejścia

Wykrywanie ludzi na obrazach i analizowanie ich zachowania w filmach wideo jest obecnie przedmiotem wielu badań z zakresu wizji komputerowej. W wielu przypadkach systemy te są stosowane do monitorowania i ochrony życia człowieka [41, 65, 90, 131]. Istnieje wiele systemów monitorowania z rozpoznawaniem potencjalnie niebezpiecznych zdarzeń (np. udaru lub upadku) w celu ochrony osób starszych za pomocą kamer RGB i RGB-D [17, 74]. Ponadto kamery mogą być stosowane do ochrony pasażerów na lotniskach poprzez automatyczne wykrywanie pozostawionych toreb, które mogą być bezpośrednim źródłem zagrożenia [122].

Dlatego wiele systemów próbuje analizować rejestrowaną przez kamery scenę w celu wykrycia i sklasyfikowania działań ludzi. Jest to złożony problem ze względu na dużą liczbę możliwych kombinacji postawy ciała, parametrów ciała (takich jak wzrost, waga itp.), odzieży i warunków środowiskowych, w których rejestrowane jest nagranie. Wiele badań [9, 19, 83] poświęcono wykrywaniu pieszych w scenach ruchu drogowego, co jest ważną ścieżką w rozwoju autonomicznych pojazdów.

Rozpoznawanie ludzkich działań odgrywa kluczową rolę w analizowaniu sportowców i graczy w wielu dyscyplinach sportowych [45]. W pełni automatyczne rozpoznawanie jest nadal otwartym problemem [2, 71, 76, 119], opisują różne podejścia, ale nadal są w stanie sklasyfikować tylko niewielką część interesujących zachowań. Podobny problem badawczy można znaleźć w wykrywaniu obiektów, gdzie badania koncentrują się na wykrywaniu obiektów reprezentujących określoną klasę obiektów [13, 52]. Obecny stan tych systemów jest daleki od uniwersalnego rozwiązania z możliwością zrozumienia kontekstu większości scen w prawdziwym życiu.

Również w sporcie kamery są niezbędne. Dzięki nim wszystkie nagrania na żywo mogą być transmitowane do szerszej publiczności. Co więcej, obecne zaawansowane komputerowe systemy wizyjne śledzą graczy, analizują ich ruchy i generują raporty dla reporterów, trenerów i innych osób w celu dalszej analizy [108]. Systemy śledzenia są stosowane do automatycznego wykrywania interesujących obszarów (np. miejscu na boisku, w którym dzieje się akcja). W obecnym stanie rozwiązania te mogą być również stosowane do tworzenia baz danych dla drużyn opisujących ich ruch i zachowanie na boisku. Jest to złożenie obszernej analizy zakończonej znalezieniem mocnych i słabych stron zarówno drużyny, jak i każdego indywidualnego zawodnika [61]. Wiedza zebrana przez obliczenia wizyjne jest następnie stosowana w podejściu statystycznym i uczeniu maszynowym.

Niemniej jednak, kamery w wielu wydarzeniach sportowych odgrywają kluczową rolę. Obecne systemy stosują kilka kamer na stadionach do analizy ruchów zawodników. Dane z kamer są stosowane do analizy zawodników w piłce nożnej [91], tenisie [102] i innych popularnych sportach [12, 40, 101]. Niektóre systemy są bardzo złożone i stosują ponad dziesięć

kamer, które rejestrują obraz w częstotliwości 340 klatek na sekundę, aby wygenerować modele 3D trajektorii piłki w krykiecie [108].

Zastosowanie podobnych technik w sportach walki nie jest tak zaawansowane, jak w poprzednich przykładach. Głównym tego powodem jest brak publicznie dostępnych oznaczonych danych, jak w przypadku innych sportów [42, 62, 79]. Jest to również spowodowane znacznie mniejszą popularnością boksu w porównaniu np. do piłki nożnej.

5.4.2. Metodologia

Głównym celem tego etapu prac jest zaproponowanie podejścia do wykrywania ciosów w boksie olimpijskim. W tym celu zastosowano jedną statyczną kamerę umieszczoną przed ringiem bokserskim. Jak zostało to omówione dalej, niektóre ograniczenia obecnego podejścia można zniwelować za pomocą zastosowania wielu kamer. Zaproponowane podejście łączy w sobie następujące dwa moduły:

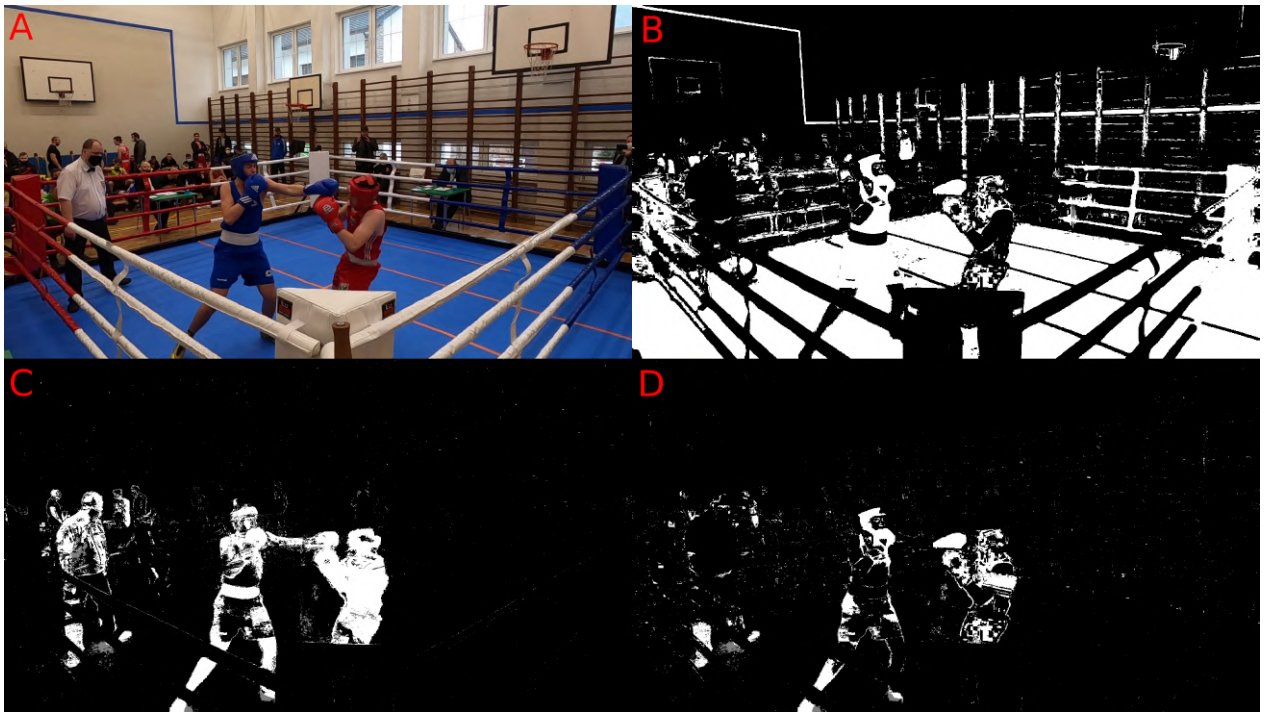
- wykrywanie starć - podejście oparte na wykrywaniu bokserów (podrozdział 5.1 oraz [99]) i mierzeniu odległości (podrozdział 5.2 oraz [100]) między nimi przy użyciu odległości euklidesowej w celu wykrycia możliwych sytuacji z ciosami.
- Wykrywanie ciosów - podejście stosujące sieci neuronowe do klasyfikowania klatek i wykrywania rzeczywistego momentu uderzenia.

Głównym celem wykrywania starć jest zmniejszenie ilości danych przed procesem etykietowania danych (co zostało wyjaśnione w rozdziale 4). Redukcja danych ma również pozytywny wpływ na przetwarzanie nowych filmów; można to uznać za filtr, który redukuje momenty, w których bokserzy stoją daleko od siebie bez szansy na ciosy. Zmniejsza to czas przetwarzania i moc obliczeniową potrzebną do faktycznego wykrywania ciosów, a także zmniejsza liczbę fałszywie pozytywnych przypadków. Dokładna metoda wykrywania starć została opisana w artykule [100] oraz w podrozdziale 5.2.

W drugim kroku zaproponowanego systemu używane są tylko klatki z bokserami znajdującymi się blisko siebie. Wykrywanie ciosów to klasyfikacja binarna z dwiema klasami („punch” i „not punch”). Do klasyfikacji stosowane jest podejście oparte o sieci neuronowe. Zastosowano dobrze znaną architekturę sieci neuronowych CNN w dziedzinie klasyfikacji obrazów. Budowa struktury sieci CNN została zainspirowana biologicznymi mechanizmami interpretacji obrazu przez ludzkie oko [21].

Wykrywanie ciosów jest jednym z problemów wizji komputerowej, w którym algorytmy muszą nauczyć się wykrywać bardzo małe obiekty na obrazie. Przyjęto założenie, że klasyfikacja bez wstępnego przetwarzania nie będzie działać poprawnie. Podstawą tego założenia były wcześniejsze prace [4, 27, 50, 121], w których dostosowanie obrazu wejściowego skutkowało lepszymi wynikami klasyfikacji. Potwierdzenie tej hipotezy zastosowanej do wykrywania ciosów zostanie zweryfikowane przez wydajność klasyfikacji. W podrozdziale zaproponowano różne metody manipulowania obrazami wejściowymi przed rozpoczęciem procesu uczenia klasyfikatora w celu sprawdzenia ich wpływu na wydajności klasyfikacji.

Zostały zaproponowane i przetestowane cztery podejścia, z których trzy zostały stworzone w celu wyodrębnienia ROI (ang. Region of Interest). Przedmiotem badań są następujące podejścia:



Rysunek 5.7: Wizualizacja oryginalnego obrazu wraz z proponowanymi metodami manipulacji obrazem

- oryginalny obraz (podejście 1) - algorytm otrzymuje oryginalny obraz bez żadnych przekształceń.
- Ekstrakcja kolorów (podejście 2) - w oparciu o reprezentację modelu obrazu HSV [53], wyodrębnione zostają niebieskie i czerwone kolory ubioru bokserów. Kolory niektórych elementów ubioru są regulowane przez zasady boksu olimpijskiego i są stałe w każdej walce (podrozdział 3.3.1).
- Odejmowanie tła (podejście 3) - algorytm usuwa statyczne obiekty i elementy z wideo, pozostawiając tylko obiekty w ruchu. Algorytm bazuje na rozkładzie gaussowskim [135].
- Metoda hybrydowa (podejście 4) - połączenie dwóch poprzednich podejść: usunięcie kolorów innych niż niebieski lub czerwony, a następnie usunięcie elementów statycznych, takich jak podłoga lub narożniki ringu bokserskiego.

Aby przetestować postawioną hipotezę, stworzono model uczenia maszynowego klasyfikacji binarnej stosujący sieć neuronową CNN. W celu statystycznego potwierdzenia wydajności klasyfikacji, każde podejście zostało przeszkolone i ocenione trzydzieści razy przy użyciu losowego podzbioru z całego zbioru danych. Dlatego w 5.4.3 przedstawiono mediany obliczonych wskaźników wydajności: dokładność (wzór (2.19)), zrównoważona dokładność (wzór (2.22)), precyzja (wzór (2.20)) i pokrycie (wzór (2.21)). Ponieważ liczba klatek wideo z uderzeniami jest o kilka razy mniejsza niż klatek bez uderzeń (dane są wysoce niezrównoważone), zrównoważona dokładność powinna zapewnić lepszą ocenę wydajności.

5.4.3. Eksperymenty

Eksperymenty przeprowadzone zostały na komputerze z 16-rdzeniowym procesorem Intel Core i9-11900K @ 3,50 GHz, 64 GB pamięci RAM i kartą graficzną Nvidia Geforce GTX 1080Ti, na systemie operacyjnym Ubuntu.

W sumie zostały wytrenowane cztery klasyfikatory przy użyciu zestawów danych przekształconych zgodnie z przedstawioną metodologią 5.4.2. Każdy model na warstwie wejściowej uzyskał obrazy w rozdzielczości 180x180 px z kanałem koloru, dodatkowo klasyfikatory zostały wytrenowane z zastosowaniem optymalizatora Adam, a entropia krzyżowa została zastosowana jako funkcja straty. Model na warstwie wyjściowej posiadał jeden neuron określający przynależność przekazanego obrazu wejściowego do jednej z dwóch klas: „punch" lub „not punch". Dane, na których modele były trenowane zostały opisane w rozdziale 4.

Model dla każdego podejścia został wytrenowany i oceniony trzydzieści razy przy użyciu losowego podzbioru z całego zbioru danych w celu statystycznego potwierdzenia skuteczności klasyfikacji. Dlatego tabela 5.6 zawiera mediany obliczonych wskaźników.

Zgodnie z oczekiwaniami, tabela 5.6 i rysunek 5.8 zawierają wyniki, na podstawie których można wnioskować, że podejście do wykrywania uderzeń z oryginalnego obrazu bez żadnych kroków wstępnego przetwarzania nie jest optymalne. Model jest bardzo niestabilny i wielokrotnie klasyfikator generuje stały, jednoklasowy wynik. Potwierdzają to również wyniki pokrycia przedstawione na rysunku 5.8. Problem z tendencyjnością klasyfikatora do głosowania na pojedynczą klasę jest głównym wyzwaniem w pracy z niezrównoważonymi danymi. W celu wygenerowania bardziej stabilnych i wydajnych klasyfikatorów przetestowane zostały kolejne podejścia wcześniej opisywane w metodologii 5.4.2.

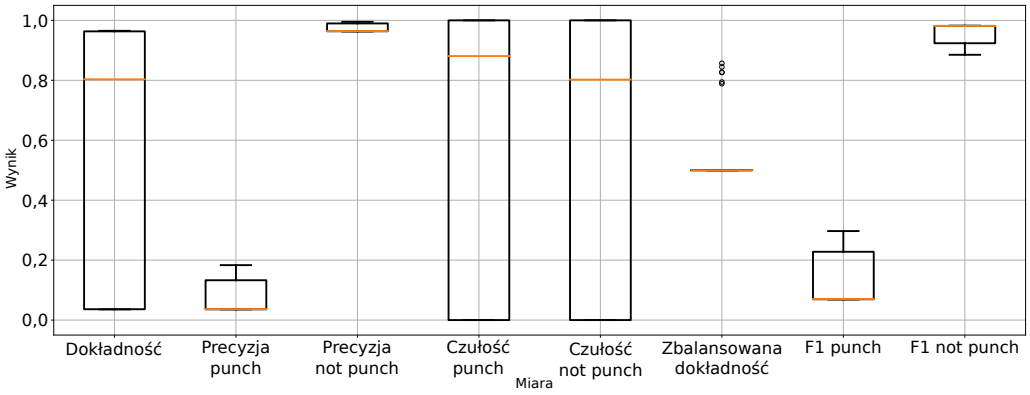
Tabela 5.6: Mediany wskaźników wydajności klasyfikacji dla wszystkich podejść

Miara	Pod. 1	Pod. 2	Pod. 3	Pod. 4
Dokładność	0,8033	0,9240	**0,9502**	0,9449
Zrównoważona dokładność	0,5000	**0,8299**	0,8257	0,7922
Precyzja „punch"	0,0364	0,2767	**0,3899**	0,3529
Precyzja „not punch"	0,9643	**0,9893**	0,9882	0,9854
Pokrycie „punch"	**0,8808**	0,7377	0,6935	0,6260
Pokrycie „not punch"	0,8023	0,9311	**0,9600**	0,9578
F1 „punch"	0,0703	0,3987	**0,4932**	0,4495
F1 „not punch"	**0,9812**	0,9594	0,9738	0,9711

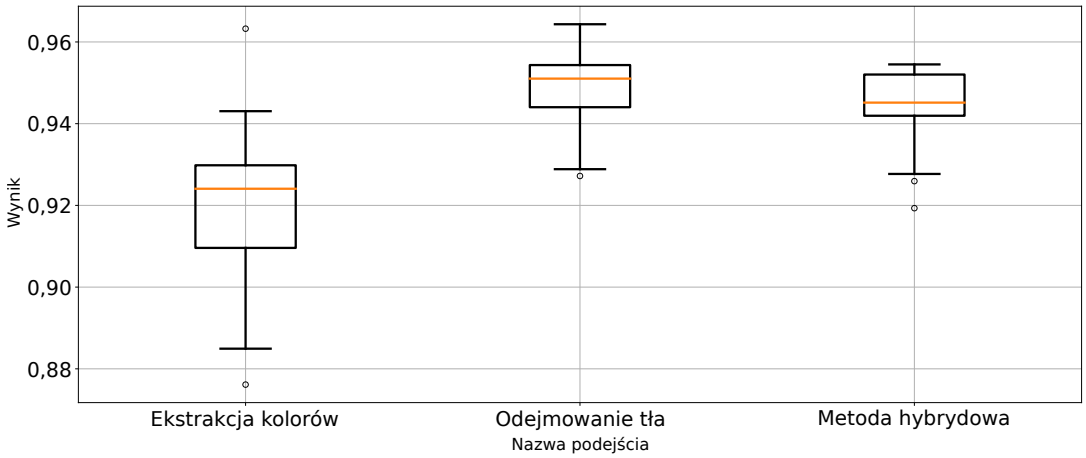
Wskaźniki wydajności klasyfikatora stosującego nieprzetworzone obrazy przedstawiono na osobnym rysunku 5.8. Traktowane są jako punkt odniesienia dla innych wskaźników wydajności przedstawionych na oddzielnych rysunkach 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16. Dodatkowo, aby poprawić czytelność, wartości odstające zostały odfiltrowane za pomocą reguły Three Sigma Rule [63] przedstawionej w równaniu (5.3). Liczba usuniętych obserwacji odstających została odnotowana w podpisie każdego rysunku.

$$\{a - 3\sigma < X < a + 3\sigma\} \tag{5.3}$$

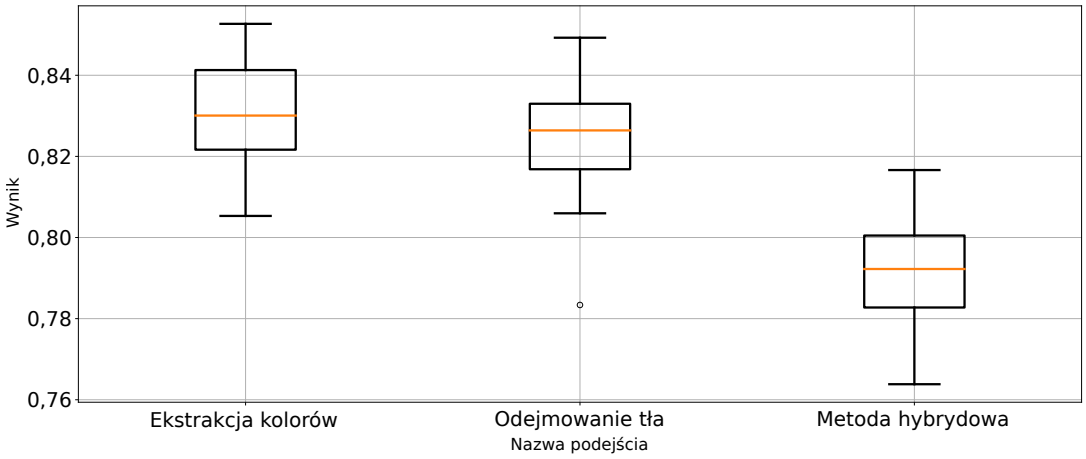
gdzie a to średnia arytmetyczna, a σ to odchylenie standardowe.



Rysunek 5.8: Wydajność klasyfikacji na oryginalnych obrazach



Rysunek 5.9: Dokładność dla trzech proponowanych podejść (usunięto 1 wartość odstającą)



Rysunek 5.10: Zrównoważona dokładność dla trzech proponowanych podejść (usunięto 4 wartości odstające)

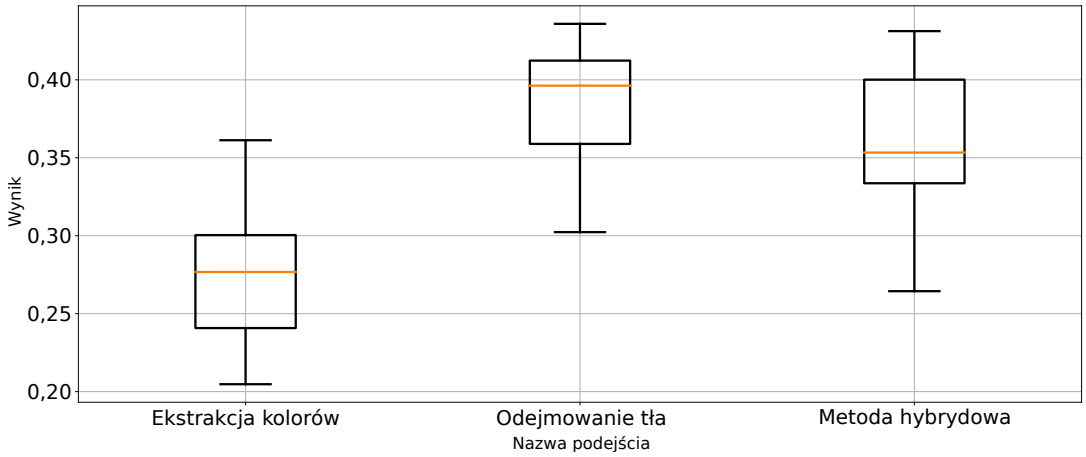
Rysunek 5.9 zawiera wykresy dokładności dla każdego z trzech testowanych podejść, najniższą medianę uzyskuje podejście stosujące odejmowanie tła, co również można zauważyć w tabeli 5.6, podejście bazujące na metodzie hybrydowej uzyskiwało zbliżone lecz gorsze wyniki. Podejście bazujące na ekstrakcji kolorów miało najniższą medianę dokładności, a sam wykres pudełkowy jest wyższy od pozostałych, co oznacza, że to podejście jest mniej stabilne względem pozostałych.

Rysunek 5.10 zawiera wykresy miary zrównoważonej dokładności, gdzie najwyższą medianę uzyskuje podejście bazujące na ekstrakcji kolorów, a same wykresy pudełkowe są podobnej wysokości. Najniższą medianę dla tej miary uzyskuje podejście hybrydowe.

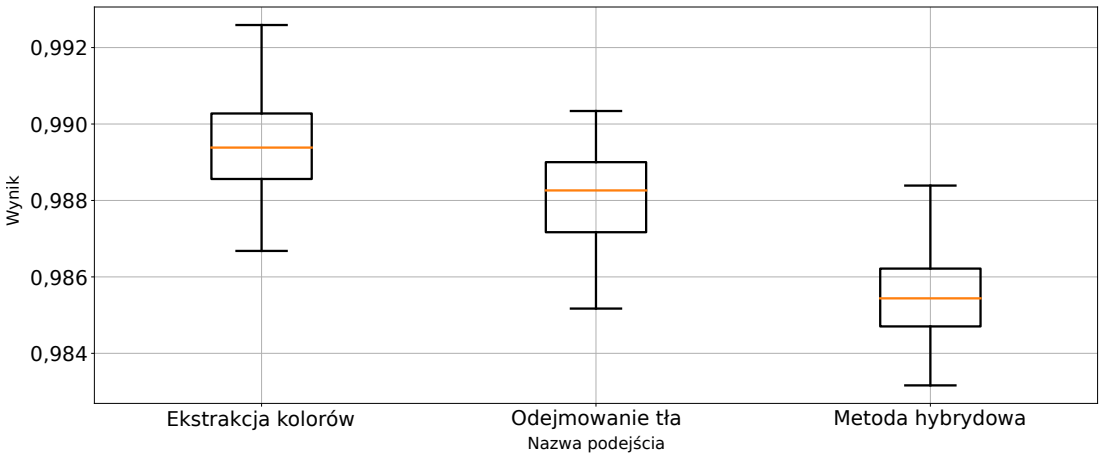
Rysunek 5.11 zawiera pudełkowe wykresy precyzji dla klasy „punch", dla tej miary najniższy wykres pudełkowy oraz najwyższą medianę uzyskało podejście bazujące na odejmowaniu tła. Dlatego właśnie to podejście powinno być stosowane w sytuacjach, w których ważne jest minimalizowanie błędnie oznaczonych klatek do klasy „punch".

Rysunek 5.12 zawiera wykresy miary precyzji dla klasy „not punch", na których można zauważyć wykresy pudełkowe podobnych wielkości oraz najwyższą medianę dla podejścia bazującego na ekstrakcji kolorów. Natomiast rysunek 5.13 zawiera wykresy miary pokrycia dla klasy „punch", które zawierają informację ile % klatek tej klasy udało się prawidłowo zaklasyfikować przy zastosowaniu każdego z podejść. W tym przypadku podejście bazujące na ekstrakcji kolorów uzyskało najwyższą medianę, a jego wielkość wykresu jest porównywalna z pozostałymi.

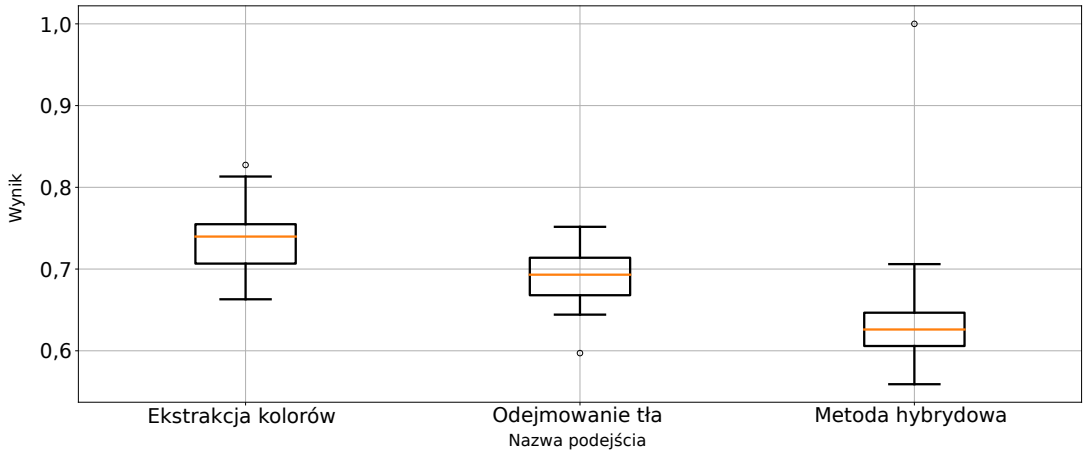
Rysunek 5.14 zawiera wykresy miary pokrycia dla klasy „not punch", na którym podejście bazujące na odejmowaniu tła uzyskało najwyższą medianę przy jednoczesnym zachowaniu najmniejszego rozmiaru wykresu. Ponadto rysunek 5.15 zawiera wykresy miary F1 dla klasy „punch", miara ta jednocześnie jest harmonijną średnią precyzji i czułości (podrozdział 2.6) dla której podejście bazujące na odejmowaniu tła uzyskuje najwyższą medianę. To samo podejście uzyskało również najwyższą medianę dla miary F1 dla klasy „not punch" co można zaobserwować na rysunku 5.16.



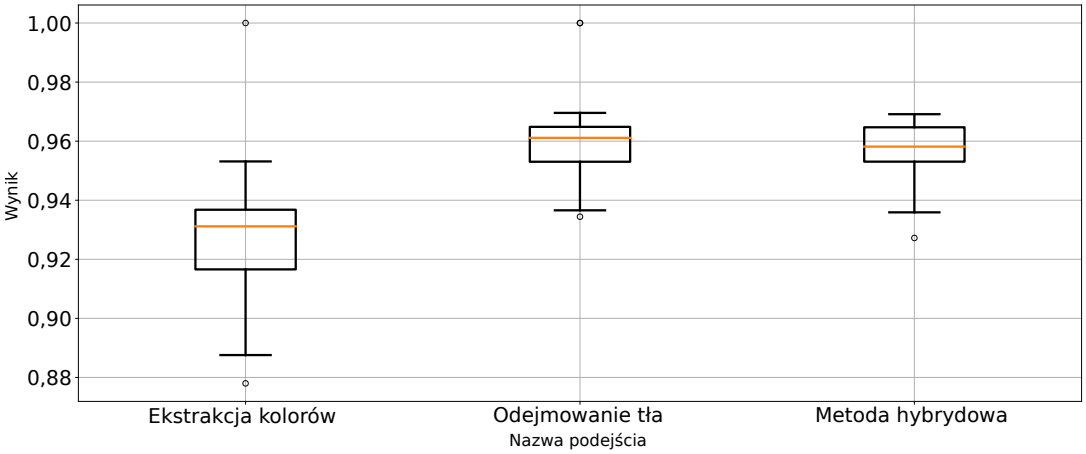
Rysunek 5.11: Precyzja dla klasy „punch” dla trzech proponowanych podejść (usunięto 1 wartość odstającą)



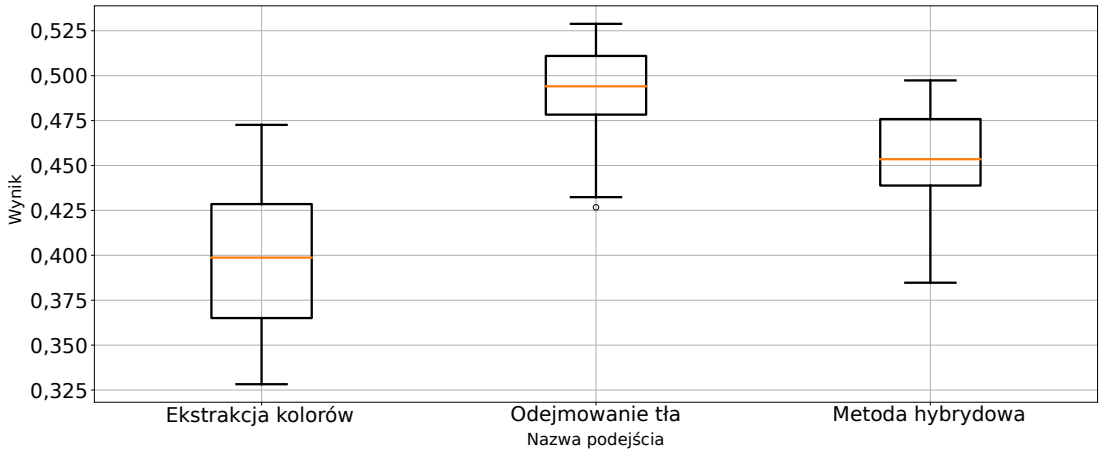
Rysunek 5.12: Precyzja dla klasy „not punch” dla trzech proponowanych podejść (usunięto 3 wartości odstające)



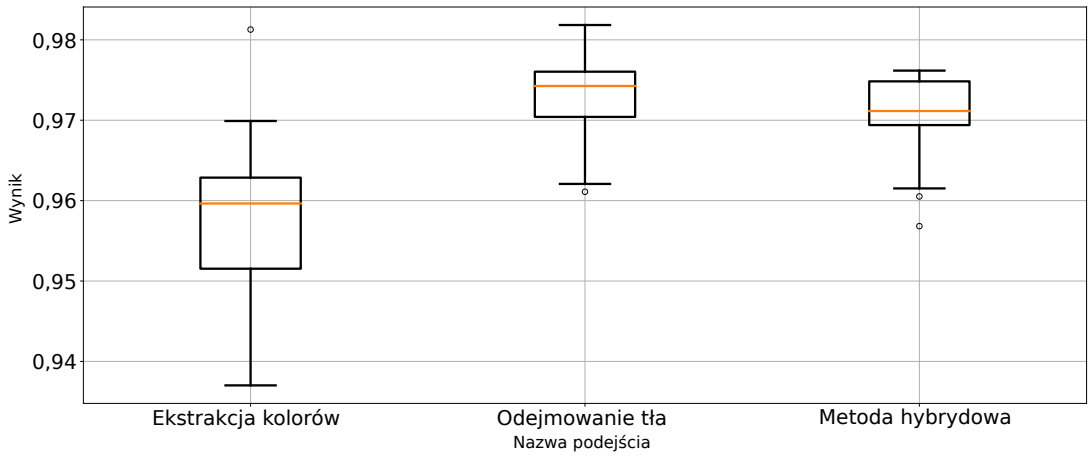
Rysunek 5.13: Pokrycie dla klasy „punch” dla trzech proponowanych podejść (usunięto 3 wartości odstające)



Rysunek 5.14: Pokrycie dla klasy „not punch” dla trzech proponowanych podejść (usunięto 1 wartość odstającą)



Rysunek 5.15: Miara $F1$ dla klasy „punch” dla trzech proponowanych podejść (usunięto 1 wartość odstającą)



Rysunek 5.16: Miara $F1$ dla klasy „not punch” dla trzech proponowanych podejść (usunięto 1 wartość odstającą)

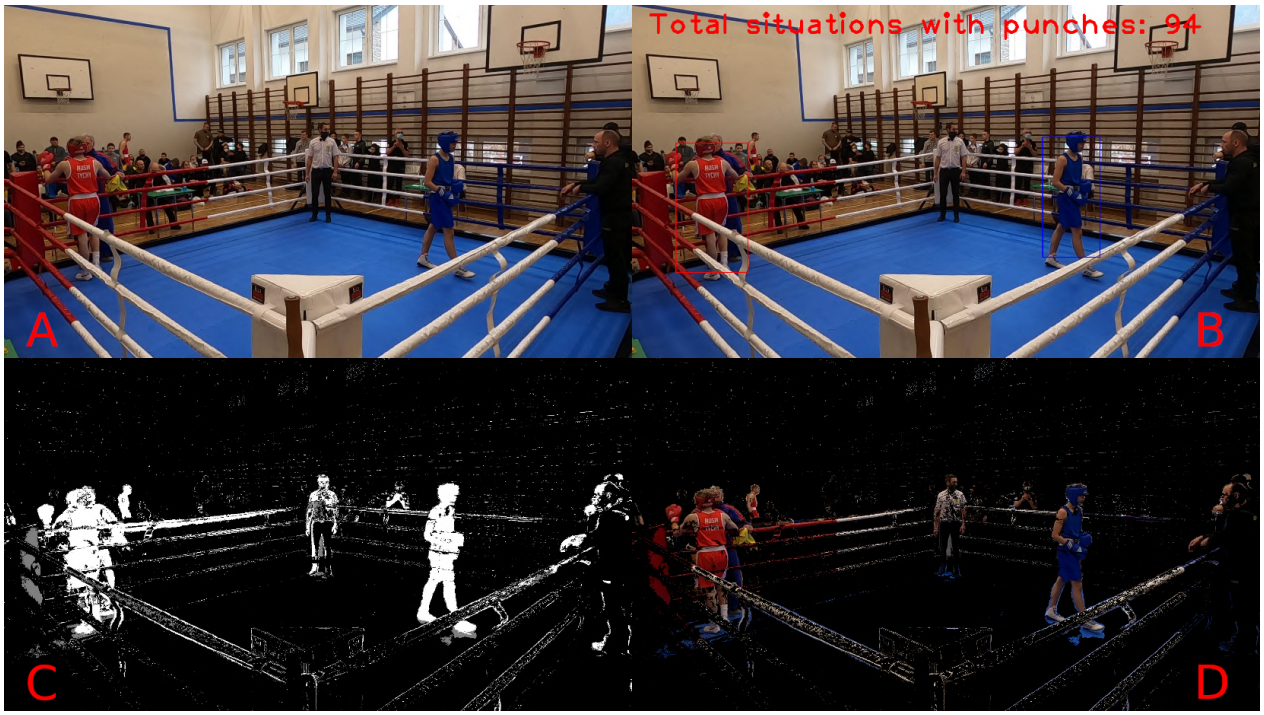
Rysunek 5.17: Przegląd całego potoku przetwarzania dla sytuacji z ciosem. A: oryginalny obraz; B: oryginalny obraz z informacją o wykrytych zdarzeniach; C: maska proponowanego podejścia 3; D: oryginalny obraz z zastosowaną maską proponowanego podejścia 3

Rysunki 5.18 i 5.17 zawierają pełną wizualizację poszczególnych kroków podczas wykrywania ciosów. Pierwszy z nich zawiera sytuację walki wręcz bez ciosów, podczas gdy drugi - wyraźnie informuje o trafieniu. Każda klatka na tych rysunkach odpowiada pojedynczemu krokowi w proponowanym systemie: od oryginalnego obrazu, poprzez maskę detekcji elementów ruchomych, aż do ostatecznej wizualizacji wykrytych etykiet na klatce. Etykieta „NEAR" napisana na żółto oznacza, że bokserzy są blisko siebie (wykryto starcie), podczas gdy etykieta „PUNCHES" oznacza wykryte ciosy między bokserami na bieżącej scenie.

Rysunki 5.18 i 5.19 zawierają stan walki, gdy nie ma kontaktu między bokserami. Na jednym z nich bokserzy są tylko blisko siebie bez aktywnej walki z ciosami, dlatego na klatce widoczna jest tylko etykieta „NEAR". W drugiej sytuacji bokserzy rozchodzą się do swoich narożników pod koniec rundy i stoją daleko od siebie, dlatego nie została napisana żadna etykieta.

Wyniki (tabela 5.6) pokazały, że podejście 3 (odejmowanie tła) osiągnęło najlepsze wyniki przy użyciu połowy metryk. Podejście 3 osiągnęło prawie taką samą zrównoważoną dokładność i precyzję dla klasy „not punch", podczas gdy podejście 1 było nieco lepsze w przypadku pokrycia dla klasy „punch" i wyniku miary $F1$ dla klasy „not punch". Co zaskakujące, najlepsze mediany wyników w tych kategoriach zostały osiągnięte przez klasyfikator bazujący na nieprzetworzonych obrazach. Prowadzi to do wniosku, że dalsze badania nad przetwarzaniem obrazu mogą zaowocować jeszcze lepszymi wynikami. Jednocześnie stabilność numeryczna klasyfikatora nieprzetworzonych obrazów jest daleka od optymalnej, a zatem - nie nadaje się do powszechnego użytku.

Rysunek 5.18: Przegląd całego potoku przetwarzania dla sytuacji walki wręcz bez ciosów. A: oryginalny obraz; B: oryginalny obraz z informacją o wykrytych zdarzeniach; C: maska proponowanego podejścia 3; D: oryginalny obraz z zastosowaną maską proponowanego podejścia 3



Rysunek 5.19: Przegląd całego procesu przetwarzania dla sytuacji bez kontaktu. A: oryginalny obraz; B: oryginalny obraz z informacją o wykrytych zdarzeniach; C: maska proponowanego podejścia 3; D: oryginalny obraz z zastosowaną maską proponowanego podejścia 3

Dlatego można uznać, że podejście 3 ma najlepszy wpływ na wydajność klasyfikacji. Jest to dość zaskakujące, ponieważ podejście 4 było najbardziej wyrafinowane w celu usunięcia sędziego ze sceny, co okazało się skuteczne (rysunek 5.7). W tym celu zostały wykonywane dodatkowe kroki przetwarzania, które powinny poprawić wydajność. Na podstawie wyników z tabeli 5.6 Podejście 4 (metoda hybrydowa) jest nieco gorsze, a czasami równie dobre niż najlepsze podejście 3. Warto zauważyć, że podejście 4 jest rozszerzeniem podejścia 2, które skutecznie poprawia wyniki metryk. Można zatem dojść do wniosku, że ekstrakcja kolorów usuwa zbyt wiele cennych informacji ze sceny co w konsekwencji skutkuje pogorszeniem się wyników.

5.4.4. Analiza statystyczna

Wyniki eksperymentalne proponowanych podejść zostały porównane przy użyciu nieparametrycznego testu statystycznego, tj. testu Friedmana [25, 44] dla $\alpha = 0,05$. Parametry testu Friedmana przedstawiono w tabeli 5.7. Ta sama tabela zawiera średnie wartości rang dla porównywanych podejść. Wyniki dla każdej z analizowanych miar jakości klasyfikacji są stosowane do testów statystycznych.

Najwyższą rangę (1,625) uzyskano dla podejścia 3 i jest ono jednocześnie krytycznie lepsze od podejścia 1 (różnica w randze 1,6250, przy 5-procentowej różnicy krytycznej wynoszącej 1,2274). Jednocześnie jest to jedyna krytyczna różnica między wszystkimi podejściami. Ponieważ żadne z podejść nie jest krytycznie gorsze od wszystkich innych podejść, nie przeprowadza się drugiej rundy analizy statystycznej.

Bardzo podobne wartości uzyskano dla podejścia 2 i podejścia 4, odpowiednio 2,5000 i 2,6250. Są to niższe rangi niż w przypadku podejścia 3 (o 0,875 i 1,000), ale w tym przypadku nie ma krytycznej różnicy między podejściami.

Podsumowując, w wyniku analizy statystycznej podejście 3 okazuje się statystycznie najlepsze, jest lepsze (ale nie krytycznie lepsze) niż podejścia 2 i 4 i jest krytycznie lepsze niż podejście 1. Z kolei podejścia 2 i 4 są lepsze niż podejście 1.

Tabela 5.7: Wyniki testu Friedmana i średnie rangi.

	Wartości
N	8
Chi-kwadrat	6,4500
Liczba stopni swobody	3
Wartość p jest mniejsza niż	0,0917
5% różnica krytyczna	1,2274
Średnie rangi	
Podejście 1	3,2500
Podejście 2	2,5000
Podejście 3	**1,6250**
Podejście 4	2,6250

5.4.5. Podsumowanie

Analiza uzyskanych wyników dowiodła, że wykrywanie ciosów przy użyciu jednej statycznej kamery RGB jest możliwe. W tym celu połączono dwie techniki, jedną do pomiaru odległości między bokserami w celu wykrycia starć między nimi. Druga była bardziej złożona i służyła do wykrywania uderzeń między bokserami w starciach.

Jak wykazały eksperymenty, wykrywanie uderzeń przy użyciu klasyfikatora wyszkolonego na nieprzetworzonych obrazach uzyskuje niskie wyniki. Zaproponowano więc trzy nowe podejścia, aby ułatwić klasyfikację nagranych scen. Podejście z literatury polegające na usuwaniu statycznych elementów z materiału filmowego ma najlepsze wyniki w niespełna wszystkich obliczonych metrykach jakości klasyfikacji.

Przeprowadzana została również analiza statystyczna, by potwierdzić znaczące różnice między wynikami dla poszczególnych metod wstępnego przetwarzania. Wyniki testów statystycznych jednoznacznie wskazują, że podejście bazujące na odejmowaniu tła jest najlepsze pod kątem wydajności klasyfikacji, co potwierdza jego skuteczność w kontekście analizy walk bokserskich.

Zaproponowano również działający system do analizy sceny bokserskiej, który oznacza bokserów, etykietuje klatki z wykrytymi starciami i ciosami oraz zlicza wszystkie sytuacje z ciosami między bokserami. Warto zauważyć, że system jest gotowy do automatycznego oznaczania nagranych materiałów, na przykład do automatycznego generowania krótkich klipów z walk.

Podsumowując, eksperymenty wykazały znaczący wpływ segmentacji obrazu na jakość klasyfikacji, dlatego kolejne etapy prac zostaną poświęcone budowie własnego rozwiązania do szybkiej segmentacji obrazu z kamery przed przekazaniem go do klasyfikatora.

6. Optymalizacja procesu klasyfikacji scen bokserskich z zastosowaniem segmentacji obrazu wideo

W analityce sportowej, szczególnie w kontekście boksu, klasyfikacja klatek wideo jako zawierających cios (klasa „punch") lub nie zawierających ciosu (klasa „not punch") stanowi poważne wyzwanie, które opiera się na drobnych szczegółach, które charakteryzują ten sport. Badania przedstawione w podrozdziale 5.4 mające na celu zautomatyzowanie procesu klasyfikacji, napotkały znaczące problemy związane z generalizacją modelu podczas stosowania konwolucyjnych sieci neuronowych (CNN) do klasyfikacji pojedynczych klatek z nagrań walk bokserskich. Wyzwania te wynikały przede wszystkim z niezdolności modeli do skutecznego uogólniania w szerokim zakresie ruchu i interakcji przedstawionych na klatkach, co doprowadziło do ponownego rozważenia tego podejścia.

Po dogłębnym przeglądzie literatury stwierdzono, że napotkany problem klasyfikacji jest ściśle związany z kwestią niskiego stosunku obszaru zainteresowania (ang. region of interest, ROI) do całego rejestrowanego obrazu, co jest powszechne w kilku problemach poza analityką sportową. Sedno problemu leży w fakcie, że obiekty, które określają klasę decyzyjną, takie jak wyprowadzenie ciosu, zajmują niewielką część całego obrazu, to jest mniej niż 1,5% całkowitego obszaru. Ten niewielki stosunek ROI do całego obrazu znacznie komplikuje zadanie dokładnej identyfikacji i klasyfikacji odpowiednich działań, ponieważ większość danych obrazu składa się z nieinformacyjnego tła lub niepowiązanych treści, które mogą stanowić jedynie szum dla modelu klasyfikacji.

Obserwacja ta jest zgodna z obserwacjami z innych dziedzin, takich jak obrazowanie medyczne i analiza obrazów satelitarnych, gdzie skuteczna klasyfikacja obrazów często zależy od identyfikacji małych, krytycznych dla decyzji obiektów lub cech znajdujących się w dużym polu widzenia. Wyzwanie to jest jeszcze trudniejsze w przypadku analizy klatek wideo ze względu na wymiar czasowy, w którym ciągłość i ewolucja ruchu w różnych klatkach wprowadza dodatkową złożoność.

Aby sprostać wyzwaniom związanym z niskim stosunkiem obszaru zainteresowania (ROI) do całego obrazu, przeprowadzono kompleksowy przegląd istniejących podejść, aby zrozumieć najlepsze praktyki w zakresie poprawy wydajności klasyfikacji przy tych ograniczeniach. Ponadto rozdział ten wprowadza nowatorskie podejście do segmentacji klatek wideo, które znacznie poprawia jakość klasyfikacji względem podejścia bazowego. Metoda ta, zaprojektowana specjalnie do szybkiej analizy sportowej, uwzględnia dynamikę boksu, w którym typowe są bardzo szybkie ruchy. Tradycyjne metodologie, choć skuteczne do pewnego momentu, nie spełniają wymagań przetwarzania obrazu w czasie rzeczywistym przy rejestrowaniu obrazu w dużej liczbie klatek na sekundę, które są niezbędne do uchwycenia szybkich akcji pomiędzy zawodnikami.

Zaproponowana w tym rozdziale technika segmentacji nie tylko osiąga znaczną popra-

wę wydajności klasyfikacji, ale także działa wielokrotnie szybciej niż testowane podejścia z literatury. Szybkość ta ma kluczowe znaczenie dla przetwarzania w czasie rzeczywistym, szczególnie w scenariuszach, w których używane są kamery rejestrujące obraz w dużej liczbie klatek na sekundę. Zastosowanie takich kamer jednocześnie zapewnia, że żaden krytyczny moment nie zostanie pominięty podczas rejestrowania szybkiej akcji bokserskiej. Ponadto zmniejszona złożoność obliczeniowa proponowanego podejścia sprawia, że dobrze nadaje się do wdrażania na urządzeniach końcowych, gdzie moc obliczeniowa jest ograniczona, a wydajność jest najważniejsza.

6.1. Inne podejścia

Podrozdział ten zawiera wyzwania i postępy w dziedzinie klasyfikacji i segmentacji obrazów, zwłaszcza w przypadku pracy z małymi obiektami na dużych obrazach. Przegląd rozpoczyna się od analizy wpływu niskiego stosunku obszaru zainteresowania (ROI) do całego analizowanego obrazu na wydajność klasyfikacji, dodatkowo zawiera innowacyjne rozwiązania, takie jak wyspecjalizowane architektury sieci neuronowych i techniki segmentacji w celu lepszego skupienia się na obszarach zawierających informacje determinujące klasę decyzyjną. W sekcji opisano zarówno tradycyjne, jak i nowoczesne strategie segmentacji, podkreślając ich rolę w poprawie wydajności klasyfikacji obrazu. Analiza ta stanowi podstawę do wprowadzenia nowatorskiego podejścia do segmentacji obrazu, które ma na celu uwzględnienie unikalnych cech szybkiej analityki sportowej jednocześnie przezwyciężając ograniczenia istniejących metodologii.

6.1.1. Problem z wydajnością klasyfikacji na obrazach z małymi obiektami

Istotne wyzwanie w dziedzinie klasyfikacji obrazów wynika z negatywnego wpływu niskiego stosunku ROI do całego obrazu na wydajność klasyfikacji. Wyzwanie to pojawia się, gdy krytyczne obiekty lub cechy niezbędne do podjęcia trafnych decyzji klasyfikacyjnych zajmują tylko niewielką część całej powierzchni obrazu. Taki scenariusz jest szczególnie powszechny w dziedzinach takich jak analiza obrazów medycznych lub satelitarnych, gdzie kluczowe elementy diagnostyczne lub informacyjne mogą być małe w stosunku do rozległego pola widzenia.

Aby rozwiązać ten problem, w pracy [54] została zaprezentowana sieć Zoom-In. Architektura ta została specjalnie zaprojektowana do klasyfikacji obrazów, w których obiekty informacyjne są znacznie mniejsze w porównaniu do rozmiaru całego analizowanego obrazu. Włączając hierarchiczne próbkowanie uwagi i uczenie kontrastowe, sieć Zoom-In nie tylko rozwiązuje podstawowe wyzwanie związane z niskim stosunkiem ROI do całego obrazu, ale także radzi sobie z ograniczeniami sprzętowymi związanymi z przetwarzaniem bardzo dużych obrazów przy ograniczonych zasobach pamięci. Poprzez strategiczne przydzielanie zasobów obliczeniowych do regionów obrazu, które z większym prawdopodobieństwem zawierają cechy informacyjne, metoda ta znacznie poprawia dokładność klasyfikacji przy jednoczesnym zachowaniu wydajności wykorzystania pamięci, oferując zauważalną poprawę w stosunku do tradycyjnych podejść CNN.

Podobnie, badanie [77] opisuje wpływ małych obiektów na możliwości modeli CNN do uogólniania. Badanie zawiera dogłębną analizę, na podstawie której autorzy wnioskują, że wraz ze zmniejszaniem się proporcji obszaru zawierającego obiekty informacyjne (ROI) do rozmiaru całego obrazu, sieci CNN wymagają znacznie więcej danych, aby utrzymać dokładność klasyfikacji. Ta odwrotna zależność między rozmiarem ROI, a zapotrzebowaniem na dane do skutecznego uczenia się podkreśla podstawowe ograniczenie istniejących architektur sieci neuronowych w obliczu niskiego stosunku ROI do całego obrazu.

Obie prace identyfikują i efektywnie radzą sobie z krytycznym wyzwaniem, jakim jest wydajna klasyfikacja dużych obrazów zawierających małe, lecz istotne informacyjnie obiekty. Podkreślają one negatywne konsekwencje niskiego stosunku obszaru zainteresowania (ROI) do całego obrazu dla efektywności klasyfikacji. Dzięki stosownym metodom i dogłębnej analizie, przynoszą one cenne spostrzeżenia oraz metodologie, które wzbogacają problem klasyfikacji obrazów.

6.1.2. Poprawa wydajności klasyfikacji poprzez segmentację obrazu

W problemie klasyfikacji obrazów i rozpoznawania obiektów naukowcy nieustannie badają metodologie mające na celu poprawę wydajności istniejących modeli, szczególnie w scenariuszach, w których obiekt zainteresowania zajmuje niewielką część obrazu, znaną jako niski stosunek obszaru zainteresowania (ROI) do całości obrazu. Dwa badania [21, 121] wniosły znaczący wkład w ten problem, koncentrując się na usuwaniu tła i ekstrakcji ROI, co przyczynia się do poprawy dokładności klasyfikacji.

W artykule [121] zademonstrowano metodę, w której usunięcie tła i ekstrakcja ROI zostały zastosowane do poprawy wydajności klasyfikacji liści, poprzez konwersję obrazów RGB do skali szarości, zastosowanie metody GrabCut do usuwania tła i wyodrębnienie ROI. Ekstrakcja cech z ROI zamiast całego obrazu nie tylko poprawia wskaźniki jakości klasyfikacji, ale także skraca czas wymagany do przetworzenia każdego obrazu.

Podobnie, badanie [21] wprowadza model DOG, który został zaprojektowany do usuwania tła z różnych typów obrazów (np. zwierząt, czy kwiatów) w celu poprawy wydajności rozpoznawania obiektów. W badaniu analizowano wpływ elementów tła na skuteczność sieci neuronowych, stwierdzając, że eliminacja tła może korzystnie wpłynąć na wydajność procesu rozpoznawania. Eksperymentalne wyniki potwierdzają efektywność modelu DOG w poprawie skuteczności klasyfikacji, podkreślając znaczenie usuwania tła w tym kontekście.

Wyzwania i rozwiązania omówione w wymienionych pracach pokrywają się z zidentyfikowanym problemem niskiego stosunku ROI do obrazu, który negatywnie wpływa na wydajność klasyfikacji scen w walkach bokserskich (podrozdział 5.4). Idąc śladem tych prac, w kolejnym podrozdziale przeprowadzono przegląd metodologii wstępnego przetwarzania, takich jak segmentacja obrazu, usuwanie tła oraz ekstrakcja ROI, mających na celu poprawę wydajności klasyfikacji. Celem jest zwiększenie dokładności i efektywności zadania klasyfikacyjnego poprzez skoncentrowanie się na najbardziej informatywnych fragmentach obrazu oraz ograniczenie wpływu nieistotnych informacji tła.

6.1.3. Obecne podejścia do segmentacji klatek wideo

Dwie metody odejmowania tła, które zyskały szerokie zastosowanie w popularnej bibliotece OpenCV i znacząco przyczyniły się do rozwoju dziedziny wizji komputerowej, zostały opracowane i przedstawione w publikacjach [134, 135].

Autorzy artykułu [134] koncentrują się na algorytmie adaptacyjnym, który stosuje gaussowskie modele mieszane (GMM) do szacowania modelu tła dla każdego piksela indywidualnie. Znaczącym wkładem tej metody jest jej zdolność do dynamicznego dostosowywania nie tylko parametrów mieszanin Gaussa, ale także liczby składników, w oparciu o obserwowane dane. Ta zdolność adaptacji pozwala modelowi dokładnie reprezentować złożone zmiany tła, takie jak te spowodowane przez poruszające się drzewa, zmieniające się warunki oświetleniowe lub długoterminowe zmiany w tle sceny. Wybierając liczbę komponentów dla każdego piksela w procedurze, algorytm ten może w pełni dostosować się do dynamiki sceny, znacznie poprawiając dokładność odejmowania tła, jednocześnie poprawiając wydajność przetwarzania.

Z drugiej strony, autorzy artykułu [135] zaproponowali proste i skuteczne podejście do odejmowania tła, które kontrastuje z parametryczną naturą GMM poprzez zastosowanie nieparametrycznej estymacji gęstości. Metoda ta dostosowuje rozmiar jądra dla każdego piksela, w oparciu o rozkład obserwowanych danych, w celu dokładnego modelowania tła. Taka elastyczność w doborze rozmiaru jądra, szczególnie dzięki podejściu „estymatora balonowego", pozwala modelowi skutecznie przetwarzać różne dynamiki tła. Metoda ta wyróżnia się prostotą implementacji i zdolnością do zapewnienia solidnej wydajności odejmowania tła bez konieczności skomplikowanego dostrajania parametrów.

Obie metody znacząco przyczyniły się do rozwoju technik odejmowania tła, zapewniając solidne rozwiązania adaptacyjne dla rzeczywistych zastosowań. Ich implementacja w popularnym module OpenCV sprawiła, że stały się one dostępne dla szerokiego grona użytkowników, ułatwiając rozwój zaawansowanych komputerowych systemów wizyjnych, które wymagają wydajnego i dokładnego modelowania i odejmowania tła.

6.1.4. Nowoczesne podejścia do segmentacji klatek wideo

W artykule [106] autorzy przedstawili podejście o nazwie BSUV-Net, które stanowi znaczący postęp w problemie odejmowania tła wideo, szczególnie w przypadku niewidzianych wcześniej nagrań, poprzez zastosowanie w pełni splotowej sieci neuronowej (ang. fully connected neural network, FCNN). Metoda ta wprowadza algorytm, który przewyższa najnowocześniejsze algorytmy odejmowania tła (ang. background subtraction, BGS) w kilku miarach oceny na zbiorze danych CDNet-2014 [115].

Kluczowe cechy sieci BSUV-Net obejmują jej zdolność do obsługi niewidzianych wcześniej nagrań wideo. Zdolność tę osiągnięto dzięki zastosowaniu dwóch referencyjnych klatek tła pobranych w różnych skalach czasowych, wraz z semantycznymi mapami segmentacji i nowatorską techniką rozszerzania danych zaprojektowaną w celu uwzględnienia zmian w oświetleniu sceny. Podejście to stosuje informacje semantyczne, w połączeniu z bieżącą klatką i dwiema referencyjnymi klatkami tła, aby znacznie poprawić dokładność segmentacji pierwszego planu.

Jednym z najbardziej godnych uwagi aspektów sieci BSUV-Net jest jej wydajność w różnych warunkach oświetleniowych, demonstrująca solidną zdolność do radzenia sobie ze zmianami oświetlenia, co jest częstym wyzwaniem w rozwiązaniach BGS. Architektura sieci, która

obejmowała dane wejściowe dotyczące informacji semantycznych oraz tła referencyjnego w wielu skalach czasowych, przyczyniła się do jej wysokiej wydajności.

Zdolności sieci BSUV-Net został podkreślony przez wyniki eksperymentalne na zbiorze danych CDNet-2014 [115], gdzie przewyższyła ona inne algorytmy BGS pod względem miary F1, precyzji i pokrycia. Ta wydajność była szczególnie zauważalna w kategoriach takich jak „filmy nocne", gdzie model wykazał doskonałą odporność na lokalne zmiany oświetlenia w porównaniu z resztą algorytmów w rankingu CDNet-2014[1]. Mimo to, model napotkał wyzwania w kategoriach „drgania kamery" i „dynamiczne tło", co wskazuje na obszary do dalszych badań i potencjalne ulepszenia w obsłudze filmów ze znacznym ruchem tła lub efektami rozmycia.

Podsumowując, BSUV-Net wyznacza nowy punkt odniesienia dla algorytmów BGS. Poprzez identyfikację tła w wielu skalach czasowych, segmentację semantyczną i innowacyjne rozszerzenie danych, oferuje obiecujący kierunek dla przyszłego rozwoju przetwarzania wideo i wizji komputerowej. Zaletą tego podejścia jest również publicznie dostępny kod źródłowy[2], który umożliwił porównanie go z algorytmem zaproponowanym w podrozdziale 6.3.

Boks jest jednak bardzo dynamicznym sportem, charakteryzującym się szybkimi ruchami i działaniami, które trwają ułamki sekund. Obserwacje wskazują, że niektóre ciosy mogą trwać zaledwie $\frac{1}{50}$ sekundy, co podkreśla znaczenie nagrywania filmów z minimalną liczbą klatek na sekundę (ang. frames per second, fps) wynoszącą 50 fps, aby zapewnić, że krytyczne momenty nie zostaną pominięte. Biorąc pod uwagę ten wymóg, wyzwaniem staje się przetwarzanie tych filmów z dużą liczbą fps w czasie zbliżonym do rzeczywistego bez pogorszenia wydajności klasyfikacji. Wymaga to rozwiązania, które jest mniej intensywne obliczeniowo i nie zmniejsza wydajności klasyfikacji w porównaniu z istniejącymi metodologiami. Zaspokojenie tych konkretnych potrzeb stanowi sedno niniejszej rozprawy. W tym celu wprowadzono nowatorskie podejście i opisano je w podrozdziale 6.2. Ponadto dokonano kompleksowego porównania z wcześniej opisanymi podejściami (podrozdział 6.3). Taka analiza porównawcza nie tylko podkreśla wydajność i skuteczność proponowanej metody, ale także ilustruje znaczącą lukę, którą wypełnia ona w problemie szybkiej analityki sportowej, szczególnie w boksie.

6.2. Autorskie podejście skracające czas przetwarzania

Niniejsze badanie koncentruje się na klasyfikacji poszczególnych klatek wideo z nagrania zawierającego walkę bokserską. W konfiguracji eksperymentalnej opisanej w podrozdziale 6.3, klasyfikatory binarne zostały przeszkolone do przypisywania każdej klatki do klasy „punch" lub „not punch". Klatki sklasyfikowane jako „punch" zawierają scenariusze walki, w których bokser z powodzeniem uderza przeciwnika (np. w głowę lub tułów). Natomiast klatki oznaczone jako „not punch" zawierają momenty, w których bokserzy nie angażują się aktywnie (np. znajdują się daleko od siebie), są w klinczu lub wykonują ciosy, które nie trafiają skutecznie przeciwnika (np. uderzają w blok lub rękawice).

Problem klasyfikacji zdjęć, na których znaczące obiekty o charakterze informacyjnym zajmują jedynie kilka procent powierzchni to otwarty problem naukowy, który jest zarazem

[1] http://changedetection.net (Data dostępu: 12.12.2023)
[2] https://github.com/ozantezcan/BSUV-Net-inference (Data dostępu: 12.12.2023)

przedmiotem badań w ostatnich latach [54, 55, 77]. Problem ten występuje m.in. podczas klasyfikacji zdjęć satelitarnych czy medycznych [54].

Na podstawie zebranej bazy danych walk bokserskich (proces jej budowania został opisany w rozdziale 4) można wnioskować, że powierzchnia obszaru, na którym padają ciosy bokserów zajmuje poniżej 1,5% powierzchni całej rejestrowanej sceny przez kamerę.

W celu poprawy zdolności sieci neuronowych do uogólnienia oraz podniesienia ich jakości klasyfikacji podczas pracy na obrazach, na których znaczące obiekty o charakterze informacyjnym zajmują jedynie kilka procent powierzchni, zaproponowano podejście przetwarzania wstępnego danych przed przekazaniem do klasyfikacji. Zaproponowane przetwarzanie wstępne bazuje na operacji odjęcia n-wcześniejszej klatki od obecnie przetwarzanej. Operację odjęcia jednego obrazu od drugiego opisuje wzór (1.8), a zaproponowane podejście zaprezentowane jest w algorytmie 6.2.

Aby rozwiązać problem niskiego stosunku ROI do całego obrazu, który negatywnie wpływa na wydajność klasyfikacji, zaimplementowano trzy podejścia segmentacji z literatury do wstępnego przetwarzania klatek wideo przed ich klasyfikacją. Podejścia te mają na celu wyodrębnienie najbardziej informatywnych części klatki, poprawiając w ten sposób zdolność klasyfikatora do dokładnego przypisywania klatek do klasy „punch" lub „not punch". Oprócz metod z literatury, zaproponowano również nowatorskie podejście, dlatego w podrozdziale z eksperymentami oceniono następujące podejścia:

- *original* - oryginalne klatki bez żadnych przekształceń przed klasyfikacją (podejście bazowe);
- *back_n_frames* - klatki z odjętym tłem przez proponowany Algorytm 6.2, gdzie n jest drugim parametrem Algorytmu;
- *extract_by_knn* - klatki z odjętym tłem przez algorytm oparty na K-najbliższych sąsiadach [134];
- *extract_by_mog2* - klatki z odjętym tłem przez algorytm oparty na mieszaninie gaussowskiej [135].
- $BSUV - Net$ - klatki z odjętym tłem za pomocą algorytmu $BSUV - Net$, który jest oparty na konwolucyjnych sieciach neuronowych [106].

Wszystkie podejścia do odejmowania tła z literatury zostały użyte z domyślnymi parametrami.

W eksperymentach testowane podejścia były oceniane w dwóch aspektach. Pierwszym i najważniejszym aspektem tej pracy jest czas przetwarzania oraz wykorzystanie zasobów podczas procesu segmentacji. Drugim aspektem jest wpływ procesu segmentacji na wydajność klasyfikacji. Wykorzystanie zasobów było monitorowane za pomocą wskaźników takich jak średnie wykorzystanie procesora i średnie zużycie pamięci RAM. Natomiast wydajność klasyfikacji mierzono za pomocą dokładności, zrównoważonej dokładności oraz miary F1.

Metody gromadzenia danych i etykietowania zostały szczegółowo opisane w publikacjach [98, 100] oraz w rozdziale 4. Wynikowa baza danych zawiera 312774 klatek, z czego 11345 (około 3,62%) sklasyfikowano jako „punch", a 301429 jako „not punch". Rozkład ten wskazuje na znaczną dysproporcję między tymi dwiema klasami. W związku z tym zrównoważona dokładność jest uważana za kluczową miarę do porównywania wydajności klasyfikacji

w ocenianych podejściach. Ponadto aby przyspieszyć proces uczenia i zmniejszyć dysproporcję między klasami, liczba przykładów klasy „not punch" użytych w eksperymentach została zmniejszona do 100614.

Aby wytrenować klasyfikatory, zbudowano niestandardową architekturę sieci neuronowej, która charakteryzuje się mniejszą liczbą warstw i parametrów w porównaniu do znanych modeli, takich jak ResNet czy Inception. To uproszczenie miało na celu przede wszystkim przyspieszenie procesu uczenia oraz dostosowanie się do ograniczonych zasobów obliczeniowych. Architektura ta składa się z następujących warstw (wymiary każdej warstwy wyjściowej zostały określone w nawiasach):

1. InputLayer (80, 80, 3);
2. AugmentationLayer (80, 80, 3);
3. Conv2D (78, 78, 40);
4. MaxPooling2D(39, 39, 40);
5. Conv2D (37, 37, 80);
6. MaxPooling2D(18, 18, 80);
7. Conv2D (16, 16, 80);
8. Dropout (16, 16, 80);
9. Flatten (20480);
10. Dense (80);
11. Dense (1) - warstwa wynikowa z jednym neuronem.

Architektura sieci neuronowej zastosowana w tym badaniu zawiera 1726241 parametrów, co jest znacząco mniej w porównaniu do częściej używanych, bardziej rozbudowanych sieci takich jak ResNet czy Inception. Taka redukcja znacznie przyspiesza proces uczenia.

Dodatkowo, aby złagodzić problem nadmiernego dopasowania sieci do danych podczas fazy uczenia, zastosowano techniki augmentacji (warstwa 2 w sieci neuronowej) i losowego porzucania (ang. dropout) (warstwa 9 w sieci neuronowej) [3, 113]. W warstwie augmentacji zastosowano serię losowych, ale realistycznych transformacji, które obejmowały:

1. losowe odbicie lustrzane (ang. random flip).
2. Losowe przybliżenie/oddalenie (ang. random zoom).
3. Losowy kontrast (ang. random contrast).

Algorytm 6.1: Algorytm do normalizacji wyniku odjęcia dwóch klatek od siebie

Wejście: $diff$ – wynik odjęcia dwóch klatek od siebie
Wyjście: $normalized\_diff$ – znormalizowany wynik odjęcia dwóch klatek od siebie

1 $diff[(diff <= 15)] = 0$
2 $diff[(diff > 240)] = 0$
3 $diff[diff > 15] = 255$
4 $kernel = get\_structured\_element\_3x3()$
5 $normalized\_diff = opening(diff, kernel)$ # wzór (1.13)
6 **result** $normalized\_diff$;

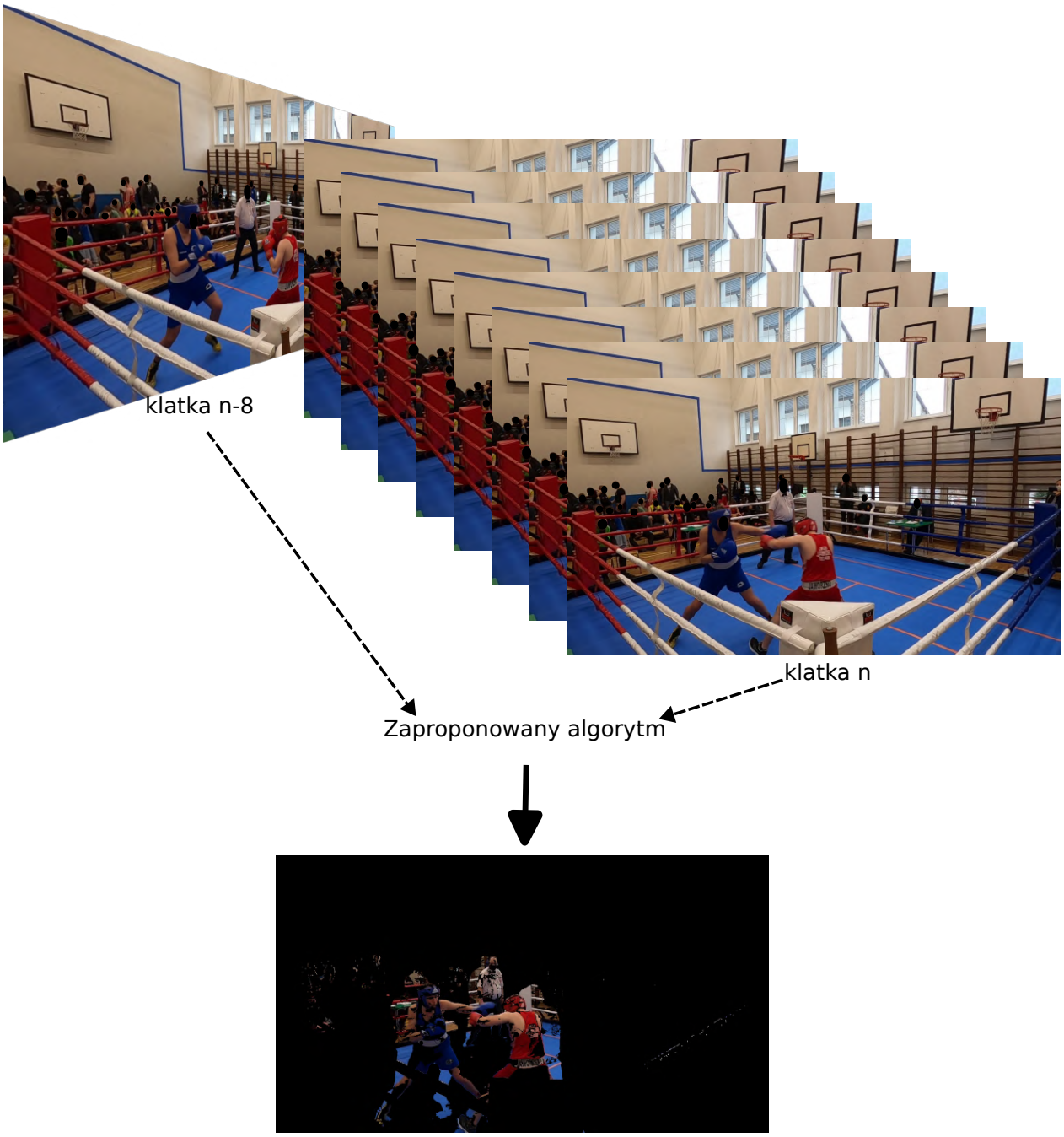
Algorytm 6.2: Zaproponowany algorytm do segmentacji klatek wideo

Wejście: $frame$ – jedna klatka z nagrania wideo
Wejście: $compare\_with\_n\_back\_frame$ – liczba określająca z o ile wcześniejszą klatką algorytm ma odjąć klatkę wejściową
Wyjście: $segmented\_image$ – klatka wideo po segmentacji

1 $gray = convert\_image\_rgb\_to\_grayscale(frame)$ # wzór (1.3)
2 $previous\_n\_frames.insert\_at\_first\_position(gray)$
3 **if** $length(previous\_n\_frames) <= compare\_with\_n\_back\_frame$ **then**
4 | $segmented\_image = frame$
5 **else**
6 | $diff = gray$ - $previous\_n\_frames[compare\_with\_n\_back\_frame]$; # wzór (1.8)
7 | $mask = Algorithm$ 6.1$(diff)$ # algorytm 6.1
8 | $segmented\_image = apply\_mask\_on\_image(frame, mask)$ # wzór (1.14)
9 | $previous\_n\_frames.remove\_last\_element()$
10 **result** $segmented\_image$;

W 1 kroku algorytmu 6.2 otrzymana klatka w kolorze jest konwertowana do odcieni szarości poprzez funkcję *convert_image_rgb_to_grayscale*, która stosuje do tej operacji wzór (1.3). W 2 kroku wykonywany jest zapis przekształconej klatki do pamięci; w 3 kroku algorytmu sprawdzane jest czy w pamięci znajduje się odpowiednia liczba klatek do operacji odjęcia, jeśli nie to zwracana jest otrzymana klatka (krok 4); jeśli pamięć zawiera już wystarczającą liczbę klatek, to algorytm przechodzi do kroków 6-9. W kroku 6 dokonuje się operacja odjęcia klatki wejściowej od n wcześniejszej klatki, operacja ta bazuje na wzorze (1.8). W kroku 7 następuje normalizacja wyniku odjęcia dwóch klatek od siebie z zastosowaniem algorytmu 6.1. W kroku 8 z zastosowaniem funkcji *apply_mask_on_image* następuje nałożenie otrzymanej różnicy, jako maski na klatkę wejściową stosując wzór (1.14). W kroku 8 zmienna *segmented_image* zawiera widoczne jedynie te obszary, w których wykryto zmianę (ruch), pozostałe obszary pozostają czarne. Krok 9 jest odpowiedzialny za czyszczenie pamięci z niepotrzebnej już ostatniej klatki; krok 11 zwraca wynik algorytmu. Graficzną wizualizację działania algorytmu oraz dostarczanych wyników zawiera rysunek 6.1.

Algorytm 6.1 odpowiedzialny za normalizację wyników odjęcia dwóch klatek od siebie (stosowany w algorytmie 6.2) w krokach 1-3 normalizuje wszystkie piksele do wartości 0 lub 255. W krokach 1 i 2 algorytmu piksele o wartości > 240 oraz $<= 15$ są zerowane, ma to na celu usunięcie wykrytych minimalnych różnic na pikselach pomiędzy klatkami, różnice te są



Rysunek 6.1: Wizualizacja proponowanego algorytmu dla n=8

spowodowane automatyczną korektą oświetlenia, która była włączona w kamerze podczas nagrywania. W kroku 3 do wszystkich pozostałych pikseli przypisywana jest wartość 255 (biały); w kroku 5 została wykonana operacja otwarcia na podstawie wzoru (1.13), która korzysta z operacji erozji oraz dylatacji (wzór (1.11) oraz (1.12)). Do operacji otwarcia niezbędnym było stworzenie elementu strukturalnego (krok 4), który odpowiada parametrowi B we wzorze (1.13); w kroku 6 następuje zwrócenie wyniku działania algorytmu.

Proponowany algorytm 6.2 został oceniony z różnymi wartościami drugiego parametru i porównany z trzema innymi podejściami z literatury w podrozdziale 6.3.

6.3. Eksperymenty

Celem eksperymentów jest porównanie proponowanego podejścia do segmentacji obrazu z innymi podejściami z literatury oraz z podejściem bazowym nie stosującym żadnych przekształceń. Testowane podejścia są oceniane pod kątem wydajności oraz wpływu na jakość klasyfikacji.

6.3.1. Konfiguracja

Badanie to zostało przeprowadzone przy użyciu języka Python w wersji 3.10 oraz następujących pakietów: OpenCV 4.7, TensorFlow 2.15 i PyTorch 2.2.2. Całość zostało uruchomione na systemie Ubuntu zasilanym przez procesor Intel Core i9-11900K i kartę graficzną Nvidia GeForce GTX 1080Ti. Narzędzia te zostały wybrane nie tylko ze względu na ich solidność i kompatybilność z przetwarzaniem obrazu w wysokiej rozdzielczości, ale także ze względu na ich wydajność w zakresie szybkości przetwarzania, co ma kluczowe znaczenie dla analizy sportowej w czasie rzeczywistym.

6.3.2. Zestaw danych i trening modelu

Baza danych, którą zastosowano w niniejszym rozdziale została uprzednio pozyskana oraz manualnie oznaczona. Szczegółowy opis procesu przygotowania tego zestawu danych został opisany w rozdziale 4.

Zastosowana baza danych zawiera 312774 klatek, ze znaczącą dysproporcją wynikającą ze stosunku liczby klatek o klasie „punch” do liczby klatek o klasie „not punch”. Aby poradzić sobie z tą dysproporcją i zwiększyć wydajność procesu uczenia, wybrano 100614 przykładów dla klasy „not punch” oraz 11345 dla klasy „punch”. Podejście to znacznie przyspieszyło proces szkolenia bez uszczerbku dla wydajności modelu. Wybrany podzbiór został strategicznie podzielony, gdzie 80% przykładów zostało przydzielonych do zestawu treningowego i 20% do zestawu testowego. Ponadto zbiór danych zastosowany do testów wydajności modeli segmentacyjnych składał się z 14331 klatek z nagrań wideo w rozdzielczości full HD.

6.3.3. Analiza wyników

Wyniki eksperymentów zostały zapisane w tabeli 6.1, 6.2 oraz rysunku 6.2. W ramach eksperymentów przetestowano zaproponowany algorytm 6.2 dla zbioru $n = \{1, 2, 3, 5, 8, 13,$

21, 34, 55} będącego podzbiorem rozwinięcia ciągu Fibonacciego. Ponadto przetestowano również 3 inne podejścia z literatury (*extract_by_knn*, *extract_by_mog2*, $BSUV - Net$), które zostały opisane w podrozdziale 6.1. Dodatkowo punkt odniesienia stanowi również podejście bazowe (*original*), które nie stosuje żadnych przekształceń na klatkach wideo. Dlatego ostatecznie tabela 6.1 zawiera wyniki dla 13 różnych przetestowanych podejść.

Tabela 6.1 zawiera wyniki dla wskaźników jakości klasyfikacji. Wiodącymi wskaźnikami dla tego problemu klasyfikacyjnego pozostaje zbalansowana dokładność oraz miara F1, głównie z powodu pracy na danych niezbilansowanych. Z tabeli można odczytać, że najlepszą zbalansowaną dokładność $(0,882)$ oraz miarę F1 dla klasy „punch" $(0,625)$ uzyskuje podejście *extract_by_knn*, natomiast najlepszy wynik dla miary F1 dla klasy „not punch" $(0,942)$ uzyskują jednocześnie podejście *extract_by_knn* oraz $BSUV - Net$. Ponadto analizując jedynie podejście bazowe (*original*) oraz zaproponowane podejścia, podejście *back_13_frames* uzyskuje w tej podgrupie najlepsze wyniki jakości klasyfikacji, to jest: $0,893$ dokładności, $0,839$ zbalansowanej dokładności, $0,583$ miary F1 dla klasy „punch" oraz $0,939$ miary F1 dla klasy „not punch".

Ponadto rysunek 6.2 zawiera mediany wskaźników precyzji i czułości dla testowanych podejść. Jak można zauważyć miara precyzji dla klasy *punch* dla proponowanego podejścia sukcesywnie rosła wraz ze wzrostem parametru n, aż do momentu kiedy parametr uzyskał wartość 13, na tej wartości następuje moment przegięcia ponieważ wyższe wartości parametru n odnotowują spadek precyzji. Podobne zachowania można zauważyć w przypadku pozostałych miar.

Dodatkowo w celach szerszej eksploracji proponowanego podejścia oraz jednoczesnego potwierdzenia, że podejście *back_13_frames* jest optymalne przeprowadzono eksperymenty testujące zaproponowany algorytm 6.2 dla zbioru $n = \{11, 12, 13, 14, 15\}$. Eksperymenty ponownie wykazały, że podejście *back_13_frames* uzyskuje w tej podgrupie najlepsze wyniki jakości klasyfikacji. Dlatego podejście *back_13_frames*, które bazuje na operacji odjęcia od obecnie przetwarzanej klatki klatkę 13 wcześniejszą, jako jedyne zostało uwzględnione w testach wydajności (tabela 6.2).

Tabela 6.2 zawiera wyniki testów wydajności 3 podejść z literatury (*extract_by_knn*, *extract_by_mog2*, $BSUV - Net$) oraz jednego proponowanego w niniejszej rozprawie (*back_13_frames*). Tabela nie zawiera podejścia bazowego (*original*), ponieważ to podejście nie stosuje żadnych przekształceń na klatkach. Jak można zauważyć proponowane podejście jest najlepsze w 2 z 3 miar, które potrzebowało $88,629$ sekund do przetworzenia partii materiału wideo, gdzie kolejne najszybsze podejście (*extract_by_knn*) było wolniejsze o $95,503$ sekund. Pod kątem średniego wykorzystania procesora CPU proponowane podejście również wypadło najlepiej i wykorzystywało jedynie $28,147\%$ dostępnego procesora, gdzie kolejne najlepsze podejście ($BSUV - Net$) wykorzystywało procesor o $20,189$ punktów procentowych więcej. Najniższe średnie wykorzystanie pamięci podręcznej przypada na podejście $BSUV - Net$, które wykorzystywało jedynie $19,089\%$ całej dostępnej pamięci podręcznej podczas przetwarzania nagrania wideo, drugim najniższy wynik $(63,760\%)$ przypada na podejście *extract_by_mog2*, proponowane podejście znajduje się na trzecim miejscu z wynikiem $69,708\%$. Warto podkreślić, że podejście $BSUV - Net$ potrzebowało aż $40731,288$ sekund $(11,314$ godzin) do przetworzenia materiału wideo składającego się z 14331 klatek. To przekłada się na analizę o szybkości $0,351$ klatki na sekundę, co może być nieakceptowalne w pewnych zastosowaniach, szczególnie w analizie obrazu rejestrowanego w wysokiej częstotliwości, jak

Tabela 6.1: Mediany wskaźników jakości klasyfikacji dla testowanych podejść - porównanie własnego podejścia, podejścia bazowego oraz podejść z literatury

Podejście	Dokładność	Zbalansowana dokładność	F1 punch	F1 not punch	Średnie rangi
original	0,846	0,500	0,511	0,924	11,500
back_1_frames	0,798	0,654	0,317	0,882	12,750
back_2_frames	0,849	0,717	0,420	0,913	11,500
back_3_frames	0,872	0,748	0,473	0,927	10,250
back_5_frames	0,880	0,797	0,530	0,931	9,000
back_8_frames	0,886	0,821	0,566	0,934	7,375
back_13_frames	0,893	0,839	0,583	0,939	5,000
back_21_frames	0,891	0,841	0,583	0,937	5,375
back_34_frames	0,883	0,853	0,579	0,932	6,875
back_55_frames	0,887	0,853	0,585	0,934	5,250
extract_by_knn	0,895	**0,882**	**0,625**	0,939	2,125
extract_by_mog2	**0,900**	0,873	0,622	**0,942**	**1,625**
BSUV-Net	0,899	0,864	0,611	**0,942**	2,375

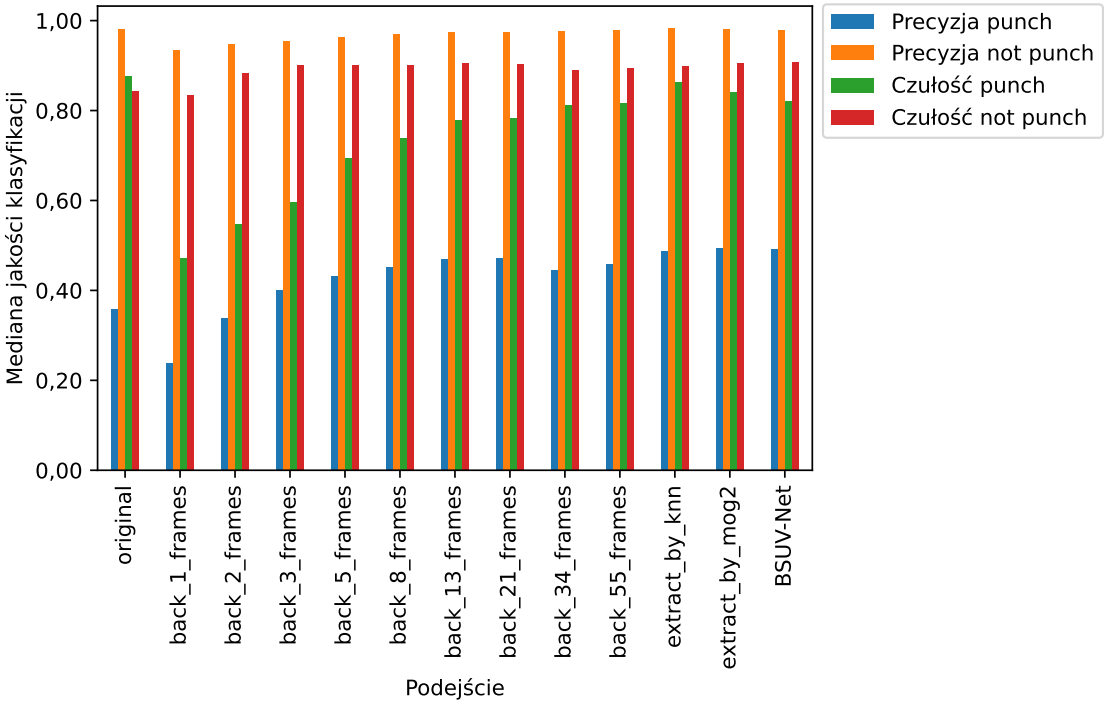
na przykład ma to miejsce w niniejszej rozprawie.

Metoda *back_13_frames* wyróżnia się minimalnym czasem przetwarzania i umiarkowanym wykorzystaniem procesora, dzięki czemu doskonale nadaje się do zastosowań w przetwarzaniu danych w czasie rzeczywistym, w których ważna jest szybkość. Jednak biorąc pod uwagę dokładność klasyfikacji, (tabela 6.1) można wnioskować, że chociaż *back_13_frames* zapewnia szybkość, nie osiąga najlepszych wskaźników wydajności pod względem zrównoważonej dokładności lub wyników miary F1. Z kolei metoda *extract_by_knn*, choć wolniejsza, wykazuje najwyższą zrównoważoną dokładność i wynik miary F1 dla przykładów o klasie „punch", wskazując na doskonałą zdolność do dokładnej klasyfikacji krytycznych klatek nagrania walki bokserskiej.

Wyniki te wskazują, że metoda *back_13_frames* jest optymalna dla scenariuszy wymagających szybkiego przetwarzania, natomiast podejście *extract_by_knn* oferuje wysoką dokładność, co jest niezbędne w zastosowaniach, w których precyzja decyzji ma kluczowe znaczenie. Ocena ta podkreśla znaczenie wyboru właściwej metody segmentacji w zależności od konkretnego zastosowania i szczegółowych wymagań. Podczas analizy wysoko dynamicznych sportów walki szybkość przetwarzania jest kluczowa, pozwala to dostarczać natychmiastową informację zwrotną dla interesariuszy, dlatego w tym przypadku podejście *back_13_frames* pozostaje najlepsze na tle innych podejść z literatury.

6.4. Dyskusja

Podrozdział zawiera krytyczną analizę implikacji wyników eksperymentalnych, badając, w jaki sposób różne techniki segmentacji klatek wideo wpływają na analitykę sportową. Analiza ta dotyczy kompromisów między wydajnością obliczeniową a dokładnością klasyfikacji,



Rysunek 6.2: Mediany wskaźników precyzji i czułości dla testowanych podejść - porównanie własnego podejścia, podejścia bazowego oraz podejść z literatury

omawiając przydatność każdej metody do konkretnych scenariuszy w środowiskach sportowych. Ponadto rozważono szerszy wpływ tych metod, nie tylko w dziedzinie sportu, ale także w innych zastosowaniach, w których analiza wideo w czasie rzeczywistym ma kluczowe znaczenie.

Wyniki podkreślają zastosowanie niezbędnego kompromisu między wydajnością obliczeniową a skutecznością klasyfikacji. Podczas gdy podejście *back_13_frames* jest liderem pod względem szybkości przetwarzania, dzięki czemu idealnie nadaje się do scenariuszy wymagających szybkiej analizy klatek, jego wydajność klasyfikacji odnotowuje nieznaczny spadek, szczególnie pod względem zrównoważonej dokładności i wyników miary F1. Z kolei metody *extract_by_knn* i *extract_by_mog2*, pomimo wolniejszego czasu przetwarzania, oferują lepszą zbalansowaną dokładność, dzięki czemu mogą być stosowane tam gdzie wymagana jest wyższa precyzja, choć z niewielkim opóźnieniem.

Biorąc pod uwagę dynamiczny i szybki charakter boksu, zdolność do szybkiego przetwarza-

Tabela 6.2: Średni czas przetwarzania i wykorzystanie zasobów dla testowanych podejść

Podejście	Czas (s)	Śr. CPU (%)	Śr. mem (%)
back_13_frames	**88,629**	**28,147**	69,708
extract_by_knn	184,132	71,190	73,904
extract_by_mog2	273,155	82,387	63,760
BSUV-Net	40731,288	48,336	**19,089**

nia klatek wideo bez znaczącej utraty dokładności klasyfikacji ma kluczowe znaczenie. Dlatego metoda *back_13_frames* stanowi wiodące podejście we wszystkich dynamicznych sportach gdzie obraz rejestrowany jest w dużej liczbie klatek na sekundę, a informacje zwrotne z analizy obrazu mają być dostarczane w czasie rzeczywistym. Zaproponowane podejście przetworzyło 14331 klatek w czasie 88, 629 sekund, co oznacza szybkość przetwarzania na poziomie 161, 697 fps. Jednoznacznie można wnioskować, że podejście jest gotowe do analizowania obrazu w czasie rzeczywistym przy kamerach rejestrujących obraz nawet w 120 klatkach na sekundę, gdzie pozostałe metody nie są już do tego przystosowane.

Podsumowując, wyniki te wskazują na konieczność dalszej optymalizacji technik segmentacji i klasyfikacji w analityce sportowej w celu poprawy zarówno szybkości, jak i dokładności, umożliwiając bardziej efektywne podejmowanie decyzji w czasie rzeczywistym i analizę wydajności. Przyszłe badania mogłyby zbadać podejścia hybrydowe, które łączą szybkość prostych metod odejmowania z cechami zwiększającymi dokładność bardziej złożonych modeli, takich jak $BSUV - Net$.

Wyniki badania mają znaczące implikacje dla problemu analityki sportowej, w szczególności w zakresie poprawy możliwości podejmowania decyzji w czasie rzeczywistym podczas wydarzeń sportowych. Wykazując różną skuteczność i dokładność technik segmentacji klatek wideo, badania te wspierają rozwój bardziej zaawansowanych narzędzi analitycznych, które można dostosować do konkretnych potrzeb różnych dyscyplin sportowych. Na przykład zdolność do szybkiego i dokładnego identyfikowania kluczowych momentów, takich jak ciosy w boksie, może mieć kluczowe znaczenie dla natychmiastowych powtórek, poprawiając uczciwość i emocje związane z tym sportem. Co więcej, techniki te można dostosować do celów treningowych, zapewniając sportowcom i trenerom natychmiastową informację zwrotną, którą można wykorzystać do dostosowania strategii i poprawy wyników podczas sesji treningowych i zawodów.

Dodatkowo, zastosowanie tych odkryć wykracza poza sport, wpływając na inne obszary wymagające analizy wideo w czasie rzeczywistym, takie jak nadzór bezpieczeństwa, czy samochody autonomiczne. W tych problemach zdolność do szybkiej i dokładnej interpretacji danych wideo może mieć kluczowe znaczenie w procesach decyzyjnych, wpływając na bezpieczeństwo i wydajność operacyjną. Praca ta stanowi fundamentalny krok w kierunku osiągnięcia wyższych prędkości przetwarzania bez uszczerbku dla dokładności zadań klasyfikacji klatek wideo, zachęcając do dalszych badań i rozwoju analizy wideo w czasie rzeczywistym w różnych obszarach.

6.5. Podsumowanie

W tym rozdziale wprowadzono i oceniono autorskie oraz inne z literatury metody segmentacji obrazu wideo w celu poprawy jakości klasyfikacji klatek nagrania walki bokserskiej do klas „punch” i „not punch”. Badania wykazały, że tradycyjne konwolucyjne sieci neuronowe (CNN) napotykają trudności w scenariuszach z niskim stosunkiem ROI (obszarem gdzie pada cios) do całego obrazu, które są powszechne w kontekście sportowym, gdzie krytyczna akcja zajmuje powierzchnię poniżej 1,5% całej klatki wideo.

Wyniki eksperymentalne podkreśliły skuteczność proponowanych technik segmentacji, w szczególności w ich zdolności do poprawy dokładności klasyfikacji poprzez skupienie się na

najbardziej istotnych sekcjach klatek wideo. Wykazano, że wprowadzone autorskie podejście znacznie przewyższa nowoczesne metody pod względem szybkości przetwarzania, dzięki czemu nadaje się do zastosowań w przetwarzaniu danych w czasie rzeczywistym. Ma to kluczowe znaczenie dla scenariuszy wymagających szybkiego podejmowania decyzji, takich jak transmisje sportowe na żywo lub w kontekstach trenerskich, w których cenne są natychmiastowe informacje zwrotne.

Przyszłe badania powinny koncentrować się na dalszym udoskonalaniu tych metod i badaniu ich zastosowania w innych dyscyplinach sportowych i zadaniach przetwarzania wideo w czasie rzeczywistym. Integracja zaawansowanych technik uczenia maszynowego i potencjał modeli hybrydowych, które łączą mocne strony różnych podejść, może prowadzić do znacznych ulepszeń w dziedzinie analityki sportowej i nie tylko. Dodatkowo, rozszerzenie tych rozwiązań na urządzenia końcowe mogłoby zminimalizować opóźnienia, jeszcze bardziej poprawiając użyteczność analizy wideo w czasie rzeczywistym w krytycznych scenariuszach decyzyjnych.

Zakończenie

Rozprawa doktorska poświęcona jest problematyce optymalizacji procesu klasyfikacji dynamicznych scen bokserskich za pomocą technik segmentacji obrazu. W rozprawie zaprezentowano autorskie podejście, które pozwala na znaczące skrócenie czasu przetwarzania danych przy jednoczesnym utrzymaniu wysokiego poziomu wydajności i stabilności klasyfikacji.

We wprowadzeniu do pracy podkreślone zostało znaczenie efektywności i dokładności w analizie wizyjnej, wskazując na potrzebę nowatorskich rozwiązań w dziedzinie przetwarzania obrazu. W pracy dodatkowo zwrócono uwagę na ograniczenia obecnych technik, które często są czasochłonne i niewystarczająco efektywne obliczeniowo przy pracy z dynamicznymi scenami i dużą ilością danych.

Rozdział 6, stanowi kluczową część rozprawy, prezentując autorskie podejście do problemu, które koncentruje się na segmentacji klatek wideo przed ich klasyfikacją. Metoda opiera się na innowacyjnym zastosowaniu operacji odjęcia n-wcześniejszej klatki, co znacząco przyspiesza proces identyfikacji istotnych zmian w rejestrowanej scenie.

Eksperymenty przeprowadzone w ramach pracy wykazały, że zaproponowane rozwiązanie redukuje czas oraz średnie wykorzystanie procesora podczas przetwarzania danych o kolejno **52%** i **42%** w odniesieniu do najlepszego testowanego podejścia z literatury. Dodatkowo zaproponowane podejście dla miary zbalansowanej dokładności odnotuje wzrost o **35** punktów procentowych względem podejścia bazowego jednocześnie nie odnotowano znacznego pogorszenia się tej miary względem innych podejść z literatury.

Efektywność obliczeniowa proponowanego rozwiązania jest szczególnie istotna w zastosowaniach wymagających szybkiego przetwarzania danych, jak analiza w czasie rzeczywistym. Znajduje również zastosowanie w analizie danych przy ograniczonych zasobach obliczeniowych, jak analiza na urządzeniach końcowych.

Tym samym osiągnięty został główny cel główny pracy, jakim było zaproponowanie rozwiązania segmentacji obrazu, które znacząco skróci czas przetwarzania danych. Do realizacji celu pracy niezbędne było również osiągnięcie celów pobocznych w tym zbudowanie bazy danych walk bokserskich, która została również opublikowana.

Podsumowując, rozprawa wprowadza znaczące ulepszenia do procesu klasyfikacji obrazów, otwierając nowe możliwości dla analizy scen sportowych i innych zastosowań wymagających szybkiej i efektywnej analizy wizyjnej. Wyniki te mają potencjalne zastosowanie nie tylko w sporcie, ale również w innych dziedzinach, gdzie kluczowe jest szybkie i dokładne przetwarzanie obrazu.

Szersze implikacje tych badań wykraczają poza boks i dotyczą innych sportów i aktywności, w których wymagana jest szybka i dokładna klasyfikacja określonych wydarzeń w strumieniach wideo. Możliwość dostosowania metod segmentacji i klasyfikacji w oparciu o dynamiczny charakter sportu i specyficzne cechy aktywności może zrewolucjonizować analitykę w czasie rzeczywistym, poprawiając zarówno wrażenia widzów, jak i rozwój strategii zespołu.

Patrząc w przyszłość, integracja bardziej zaawansowanych technik uczenia maszynowego i badanie modeli hybrydowych, które łączą różne podejścia do segmentacji, może doprowa-

dzić do znacznych postępów w tym problemie. Istnieje również obiecująca droga w badaniu rozwiązań przetwarzania na urządzeniach końcowych (ang. edge devices), aby jeszcze bardziej zmniejszyć opóźnienia, co ma kluczowe znaczenie dla podejmowania decyzji w czasie rzeczywistym w sporcie i innych aplikacjach wrażliwych na czas. Kierunki te nie tylko obiecują poprawę możliwości analityki sportowej, ale także przyczyniają się do rozwoju problemu przetwarzania wideo w czasie rzeczywistym.

Dodatkowo w przyszłości zaplanowano przetestowanie zaproponowanego algorytmu w innych problemach klasyfikacji nagrań wideo, odmiennych od walk bokserskich. Ponadto rozważane jest zaproponowanie własnej warstwy sieci neuronowej, która byłaby odpowiedzialna za segmentację kolejno przetwarzanych klatek wideo zgodnie z zaproponowanym algorytmem.

Bibliografia

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*. IEEE, aug 2017.

[2] A. Alfaro, D. Mery, and A. Soto. Human action recognition from inter-temporal dictionaries of key-sequences. In *Image and Video Technology*, pages 419–430. Springer Berlin Heidelberg, 2014.

[3] J. Ba and B. Frey. Adaptive dropout for training deep neural networks. *Advances in neural information processing systems*, 26, 2013.

[4] O. Barnich and M. V. Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, jun 2011.

[5] P. Batra, G. R. Singh, and N. Goyal. Application of adnn for background subtraction in smart surveillance system. 2023.

[6] M. Baygin, M. Karakose, A. Sarimaden, and A. Erhan. Machine vision based defect detection approach using image processing. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, sep 2017.

[7] S. K. Behendi, S. Morgan, and C. B. Fookes. Non-invasive performance measurement in combat sports. In *Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS)*, pages 3–10. Springer International Publishing, nov 2015.

[8] R. J. Brachman and T. Anand. The process of knowledge discovery in databases. advances in knowledge discovery and data mining. In *American Association for Artificial Intelligence*, pages 37–57, 1996.

[9] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. EuroCity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, aug 2019.

[10] S. Brutzer, B. Hoferlin, and G. Heidemann. Evaluation of background subtraction techniques for video surveillance. In *CVPR 2011*. IEEE, jun 2011.

[11] A. Bugeau and P. Perez. Detection and segmentation of moving objects in highly dynamic scenes. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2007.

[12] M. Buric, M. Pobar, and M. Ivasic-Kos. Object detection in sports videos. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, may 2018.

[13] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao. R-CNN for small object detection. In *Computer Vision – ACCV 2016*, pages 214–230. Springer International Publishing, 2017.

[14] C. Chen, R. Surette, and M. Shah. Automated monitoring for security camera networks: promise from computer vision labs. *Security Journal*, 34(3):389–409, feb 2020.

[15] S.-C. S. Cheung and C. Kamath. Robust background subtraction with foreground validation for urban traffic video. *EURASIP Journal on Advances in Signal Processing*, 2005(14), aug 2005.

[16] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, may 2003.

[17] C. F. Crispim, V. Bathrinarayanan, B. Fosty, A. Konig, R. Romdhane, M. Thonnat, and F. Bremond. Evaluation of a monitoring system for event recognition of older people. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, aug 2013.

[18] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection viaregion-based fully convolutional networks. *Conference: Advances in Neural Information Processing Systems*, 2016.

[19] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, apr 2012.

[20] T. D'Orazio and M. Leo. A review of vision-based systems for soccer video analysis. *Pattern Recognition*, 43(8):2911–2926, aug 2010.

[21] W. Fang, Y. Ding, F. Zhang, and V. S. Sheng. DOG: A new background removal for object recognition from images. *Neurocomputing*, 361:85–91, oct 2019.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[23] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

[24] D. A. Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[25] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[26] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. 1997.

[27] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva. Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, 35:100204, feb 2020.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014.

[29] M. Grandini, E. Bagli, and G. Visani. Metrics for multi-class classification: an overview. 2020.

[30] H. Elbehiery, A. Hefnawy, and M. Elewa. Surface defects detection for ceramic tiles usingimage processing and morphological techniques. 2007.

[31] A. Hahn, R. Helmer, T. Kelly, K. Partridge, A. Krajewski, I. Blanchonette, J. Barker, H. Bruch, M. Brydon, N. Hooke, and B. Andreass. Development of an automated scoring system for amateur boxing. *Procedia Engineering*, 2(2):3095–3101, jun 2010.

[32] S. H. Haji and A. M. Abdulazeez. Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4):2715–2743, 2021.

[33] T. Hastie, J. Friedman, and R. Tibshirani. *The Elements of Statistical Learning*. Springer New York, 2001.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[36] L. Hui and M. Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[37] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[38] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, Sept. 1999.

[39] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer US, 2021.

[40] C. T. Jeffries. Sports analytics with computer vision. 2018.

[41] W. Jia, S. Xu, Z. Liang, Y. Zhao, H. Min, S. Li, and Y. Yu. Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector. *IET Image Processing*, 15(14):3623–3637, jun 2021.

[42] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Procedings of the British Machine Vision Conference 2010*. British Machine Vision Association, 2010.

[43] G. Kang, X. Dong, L. Zheng, and Y. Yang. Patchshuffle regularization. *arXiv preprint arXiv:1707.07103*, 2017.

[44] G. K. Kanji. *100 statistical tests*. Sage, 2006.

[45] S. Kasiri, C. Fookes, S. Sridharan, and S. Morgan. Fine-grained action recognition of boxing punches from depth imagery. *Computer Vision and Image Understanding*, 159:143–153, jun 2017.

[46] S. Kasiri-Bidhendi, C. Fookes, S. Morgan, D. T. Martin, and S. Sridharan. Combat sports analytics: Boxing punch classification using overhead depthimagery. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2015.

[47] S. Kato and S. Yamagiwa. Predicting successful throwing technique in judo from factors of kumite posture based on a machine-learning approach. *Computation*, 10(10):175, sep 2022.

[48] G. Kesavaraj and S. Sukumaran. A study on classification techniques in data mining. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, July 2013.

[49] I. Khasanshin. Application of an artificial neural network to automate the measurement of kinematic characteristics of punches in boxing. *Applied Sciences*, 11(3):1223, jan 2021.

[50] C. Kim, J. Lee, T. Han, and Y.-M. Kim. A hybrid framework combining background subtraction and deep neural networks for rapid person detection. *Journal of Big Data*, 5(1), jul 2018.

[51] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2014.

[52] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019.

[53] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. Human skin detection using rgb, hsv and ycbcr color models. *arXiv preprint arXiv:1708.02694*, 2017.

[54] F. Kong and R. Henao. Efficient classification of very large images with tiny objects. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

[55] F. Kong, X.-y. Liu, and R. Henao. Quantum tensor network in machine learning: An application to tiny object classification. 2021.

[56] S. M. Kosslyn. *Image and mind*. Harvard University Press, 1980.

[57] J. Kozak. *Decision Tree and Ensemble Learning Based on Ant Colony Optimization*. Springer International Publishing, 2019.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, may 2017.

[59] D.-S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, may 2005.

[60] H. Lee, C. Wu, and H. Aghajan. Nonstationary background removal via multiple camera collaboration. In *2007 First ACM/IEEE International Conference on Distributed Smart Cameras*. IEEE, sep 2007.

[61] M. Leo, T. D'Orazio, and M. Trivedi. A multi camera system for soccer player performance evaluation. In *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. IEEE, aug 2009.

[62] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, T. D'Orazio, and A. Distante. Real-time multiview analysis of soccer matches for understanding interactions between ball and players. In *Proceedings of the 2008 international conference on Content-based image and video retrieval - CIVR '08*. ACM Press, 2008.

[63] J. K. Leonard. Theory and problems of business statstics. *McGraw-hill*, 2004.

[64] H. Li, J. Tang, S. Wu, Y. Zhang, and S. Lin. Automatic detection and analysis of player action in moving background sports video sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(3):351–364, mar 2010.

[65] J. Li, H. Liu, T. Wang, M. Jiang, S. Wang, K. Li, and X. Zhao. Safety helmet wearing detection based on image processing and machine learning. In *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, feb 2017.

[66] J. Liang. Human boxing motion prediction using neural networks. *OA Journal of Computer Networking*, 1(2):42–48, Nov. 2022.

[67] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

[68] Y. Liu, F. Wang, J. Deng, Z. Zhou, B. Sun, and H. Li. Mogface: Towards a deeper appreciation on face detection. pages 4093–4102. arXiv, 2022.

[69] H. M and S. M.N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):01–11, mar 2015.

[70] B. Mahesh. Machine learning algorithms-a review. 2020.

[71] M. R. Malgireddy, I. Inwogu, and V. Govindaraju. A temporal bayesian model for classifying, detecting and localizing activities in video sequences. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2012.

[72] N. Mirzoeff. *An introduction to visual culture*. Psychology press, 1999.

[73] I. E. Naqa and M. J. Murphy. What is machine learning? In *Machine Learning in Radiation Oncology*, pages 3–11. Springer International Publishing, 2015.

[74] B. Ni, C. D. Nguyen, and P. Moulin. RGBD-camera based get-up event detection for hospital fall prevention. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mar 2012.

[75] S. Paneru and I. Jeelani. Computer vision applications in construction: Current state, opportunities & challenges. *Automation in Construction*, 132:103940, dec 2021.

[76] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in TV shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, dec 2012.

[77] N. Pawlowski, S. Bhooshan, N. Ballas, F. Ciompi, B. Glocker, and M. Drozdzal. Needles in haystacks: On classifying tiny objects in large images. 2019.

[78] M. M. Petrou and C. Petrou. *Image processing: the fundamentals*. John Wiley & Sons, 2010.

[79] S. A. Pettersen, P. Halvorsen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, and H. K. Stensland. Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference on - MMSys '14*. ACM Press, 2014.

[80] A. Pfeuffer, K. Schulz, and K. Dietmayer. Semantic segmentation of video sequences with convolutional lstms. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1441–1447. IEEE, June 2019.

[81] E. Quinn and N. Corcoran. Automation of computer vision applications for real-time combat sports video analysis. *European Conference on the Impact of Artificial Intelligence and Robotics*, 4(1):162–171, nov 2022.

[82] R. E. W. Rafael C. Gonzalez. *Digital Image Processing, Global Edition*. Pearson, 2018.

[83] N. K. Ragesh and R. Rajesh. Pedestrian detection in automotive safety: Understanding state-of-the-art. *IEEE Access*, 7:47864–47890, 2019.

[84] A. Raid, W. Khedr, M. El-dosuky, and M. Aoud. Image restoration based on morphological operations. *International Journal of Computer Science, Engineering and Information Technology*, 4(3):9–21, jun 2014.

[85] P. Ramya and R. Rajeswari. A modified frame difference method using correlation coefficient for background subtraction. *Procedia Computer Science*, 93:478–485, 2016.

[86] V. R. Rao, M. I. Khalil, H. Li, P. Dai, and J. Lu. Decompose the sounds and pixels, recompose the events. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2144–2152, jun 2022.

[87] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[88] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, jun 2017.

[89] A. Scott, I. Uchida, N. Ding, R. Umemoto, R. Bunker, R. Kobayashi, T. Koyama, M. Onishi, Y. Kameda, and K. Fujii. Teamtrack: A dataset for multi-sport multi-object tracking in full-pitch videos. pages 3357–3366. arXiv, 2024.

[90] J. Seo, S. Han, S. Lee, and H. Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2):239–251, apr 2015.

[91] D. Setterwall. Computerised video analysis of football– technical and commercial possibilitiesfor football coaching. *Unpublished Masters Thesis. Stockholms Universitet*, 2003.

[92] S. Sharma and A. Athaiya. Activation functions in neural networks. 2017.

[93] T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, mar 2006.

[94] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[95] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, jul 2009.

[96] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Springer US, 1993.

[97] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[98] P. Stefański, T. Jach, and J. Kozak. Classification of punches in olympic boxing using static RGB cameras. In *Computational Collective Intelligence*, pages 540–551. Springer Nature Switzerland, 2023.

[99] P. Stefański, J. Kozak, and T. Jach. The problem of detecting boxers in the boxing ring. In *Recent Challenges in Intelligent Information and Database Systems: 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28-30, 2022, Proceedings*, pages 592–603. Springer, 2022.

[100] P. Stefański. Detecting clashes in boxing. *Proceedings of the 3rd Polish Conference on Artificial Intelligence, April 25-27, 2022, Gdynia, Poland*, pages 29–32, 2022.

[101] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlucke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):13–22, jan 2018.

[102] G. Sudhir, J. Lee, and A. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*. IEEE Comput. Soc, 1998.

[103] K. Suzuki. Computerized detection of lesions in diagnostic images. In *Machine Learning in Radiation Oncology*, pages 101–131. Springer International Publishing, 2015.

[104] T. Szandała. Review and comparison of commonly used activation functions for deep neural networks. pages 203–224, July 2020.

[105] R. Szeliski. *Computer vision: algorithms and applications*. Springer London, 2011.

[106] M. O. Tezcan, P. Ishwar, and J. Konrad. Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2774–2783. arXiv, 2020.

[107] G. Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2(2-3):117–132, oct 2007.

[108] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, jun 2017.

[109] G. J. Tu, H. Karstoft, L. J. Pedersen, and E. Jørgensen. Segmentation of sows in farrowing pens. *IET Image Processing*, 8(1):56–68, jan 2014.

[110] L. Unzueta, M. Nieto, A. Cortes, J. Barandiaran, O. Otaegui, and P. Sanchez. Adaptive multicue background subtraction for robust vehicle counting and classification. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):527–540, jun 2012.

[111] M. K. Vasić and V. Papić. Multimodel deep learning for person detection in aerial images. *Electronics*, 9(9):1459, sep 2020.

[112] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. arXiv, December 2015.

[113] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

[114] D. A. Wang, C. M. S. Strauss, J. M. Springer, A. Thresher, H. Pritchard, and G. T. Kenyon. Sparse mp4. In *2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*. IEEE, Mar. 2020.

[115] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 387–394. IEEE, June 2014.

[116] N. Wattanamongkhol, P. Kumhom, and K. Chamnongthai. A method of glove tracking for amateur boxing refereeing. In *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.* IEEE, 2006.

[117] S. M. Weiss and C. A. Kulikowski. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., 1991.

[118] M. T. O. Worsey, H. G. Espinosa, J. B. Shepherd, and D. V. Thiel. An evaluation of wearable inertial sensor configuration and supervised machine learning models for automatic punch classification in boxing. *IoT*, 1(2):360–381, nov 2020.

[119] J. Wu, F. Chen, and D. Hu. Human interaction recognition by spatial structure models. In *Lecture Notes in Computer Science*, pages 216–222. Springer Berlin Heidelberg, 2013.

[120] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, sep 2015.

[121] Y.-J. Wu, , C.-M. Tsai, and F. Shih. Improving leaf classification rate via background removal and ROI extraction. *Journal of Image and Graphics*, 4(2):93–98, 2016.

[122] Z. Wu and R. J. Radke. Real-time airport security checkpoint surveillance using a camera network. In *CVPR 2011 WORKSHOPS*. IEEE, jun 2011.

[123] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian. Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4753–4762, 2016.

[124] X. Ye, B. Shi, M. Li, Q. Fan, X. Qi, X. Liu, S. Zhao, L. Jiang, X. Zhang, K. Fu, L. Qu, and M. Tian. All-textile sensors for boxing punch force and velocity detection. *Nano Energy*, 97:107114, jun 2022.

[125] A. Yilmaz, O. Javed, and M. Shah. Object tracking. *ACM Computing Surveys*, 38(4):13, dec 2006.

[126] M. D. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*, 2013.

[127] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, Feb. 2021.

[128] J. Zhang and C. H. Chen. Moving objects detection and segmentation in dynamic video backgrounds. In *2007 IEEE Conference on Technologies for Homeland Security*. IEEE, may 2007.

[129] Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[130] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, apr 2020.

[131] F. Zhou, H. Zhao, and Z. Nie. Safety helmet detection based on YOLOv5. In *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*. IEEE, jan 2021.

[132] Y. Zhou, X. Wang, M. Zhang, J. Zhu, R. Zheng, and Q. Wu. Mpce: A maximum probability based cross entropy loss function for neural network classification. *IEEE Access*, 7:146331–146341, 2019.

[133] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, and H. Yao. Event tactic analysis based on broadcast sports video. *IEEE Transactions on Multimedia*, 11(1):49–67, jan 2009.

[134] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* IEEE, 2004.

[135] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, May 2006.

[136] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, Oct. 2019.

Spis rysunków

Spis tabel