EDA REPORT

Autor: Dominik Kilian

Spis treści

1. Wstęp

- **Cel zbioru i relacja z projektem dyplomowym:**
  Zbiór danych został przygotowany na potrzeby projektu dyplomowego dotyczącego detekcji osób (Human Detection) w systemach monitoringu wizyjnego (CCTV). Celem jest wytrenowanie modelu zdolnego do identyfikacji sylwetek ludzkich w zróżnicowanych warunkach obserwacyjnych.
- **Uzasadnienie przeprowadzenia EDA:**
  Eksploracyjna analiza danych jest niezbędna, aby ocenić jakość techniczną obrazów, zidentyfikować błędy w adnotacjach, zrozumieć charakterystykę warunków oświetleniowych oraz wykryć ewentualne anomalie. Zrozumienie i optymalizacja zbioru danych przed fazą modelowania pozwala na uzyskanie modelu lepszej jakości

2. Opis zbioru danych

Dane zostały pobrane z platformy (https://roboflow.com/). Oparte są na zdjęciach z systemów monitoringu CCTV. Główny zbiór danych został zsyntezowany z 3 zbiorów:

1. https://universe.roboflow.com/project-d4kos/human-cctv/dataset/1
2. https://universe.roboflow.com/heytest/sourced-human-detection-cctv
3. https://universe.roboflow.com/dogan-cwk30/human-cctv-igzff

Łączna liczba zdjęć w zbiorze to **11 249**, zdjęcia są w formacie JPG.

Dane są udostępnione na licencji **Creative Commons Attribution 4.0 International (CC BY 4.0)**. Jest to jedna z najbardziej otwartych licencji, pozwalająca na szerokie wykorzystanie zasobów w celach naukowych lub komercyjnych.

3. Metodologia

Proces łączenia zbiorów danych przeprowadzono z wykorzystaniem autorskiego skryptu w języku Python. W trakcie syntezy dokonano wstępnej eliminacji duplikujących się obrazów, które występowały w różnych zestawach źródłowych.

Analiza EDA została przeprowadzona w środowisku programistycznym PyCharm przy użyciu narzędzia Jupyter Notebook.

Wykorzystane biblioteki:

- **OpenCV** (https://pypi.org/project/opencv-python/)
- **pandas** (https://pypi.org/project/pandas/)
- **Matplotlib** (https://pypi.org/project/matplotlib/)
- **Numpy** (https://pypi.org/project/numpy/)
- **seaborn** (https://pypi.org/project/seaborn/)
- **Pillow** (https://pypi.org/project/pillow/)

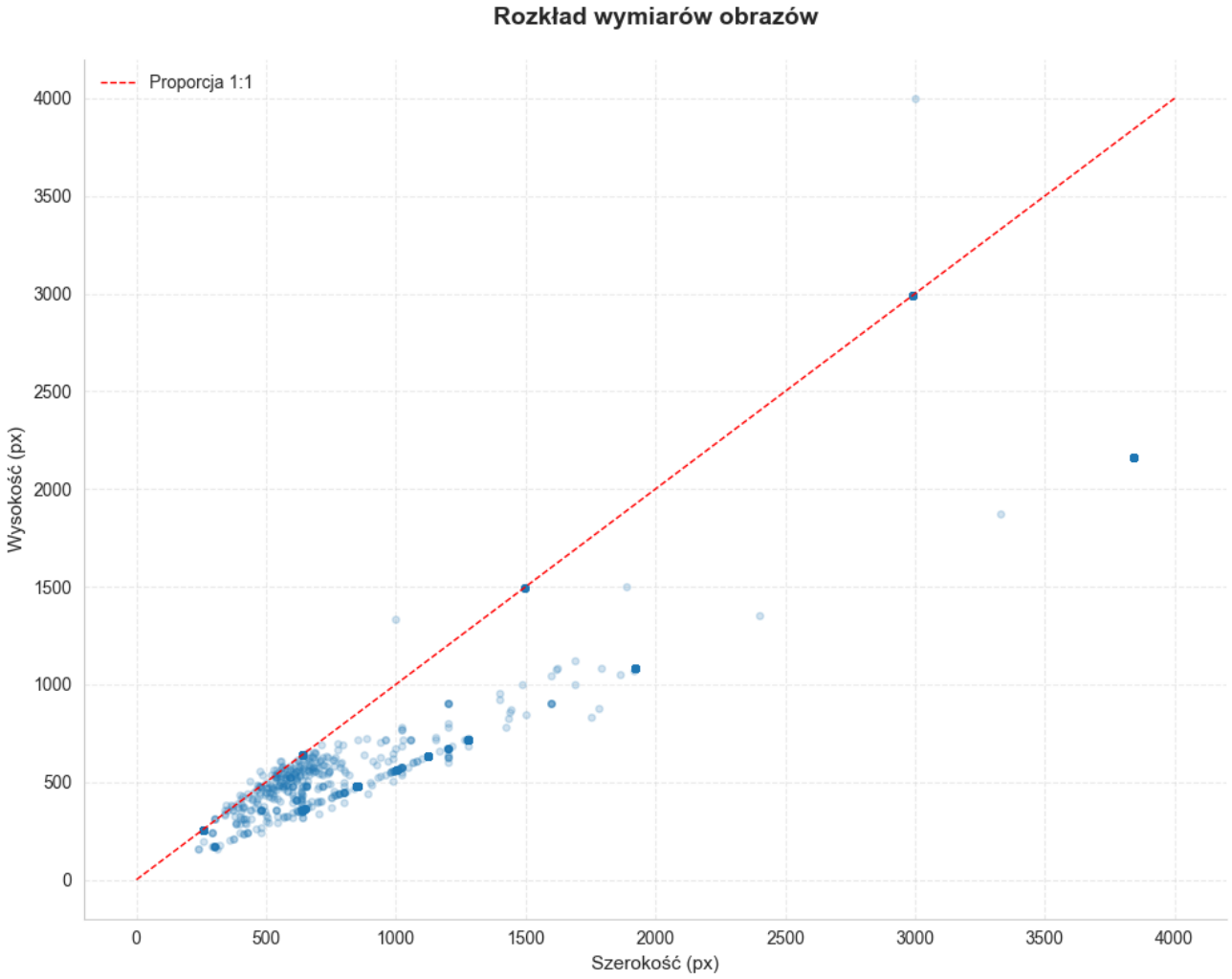
4. Wyniki analizy

Uszkodzone dane

Brak obrazów uszkodzonych lub o niepoprawnej strukturze pliku.

Rozmiary obrazów

Wykres rozkładu wymiarów obrazów pokazuje, że większość próbek posiada orientację poziomą (szerokość większa niż wysokość), co jest typowe dla nagrań z kamer monitoringu. Widoczne są wyraźne skupiska punktów odpowiadające dominującym rozdzielczościom. Widać również pojedyncze wartości odstające, reprezentujące obrazy o bardzo wysokich rozdzielczościach, wykraczających poza standardowy zbiór



5 Najczęściej występujących rozdzielczości:

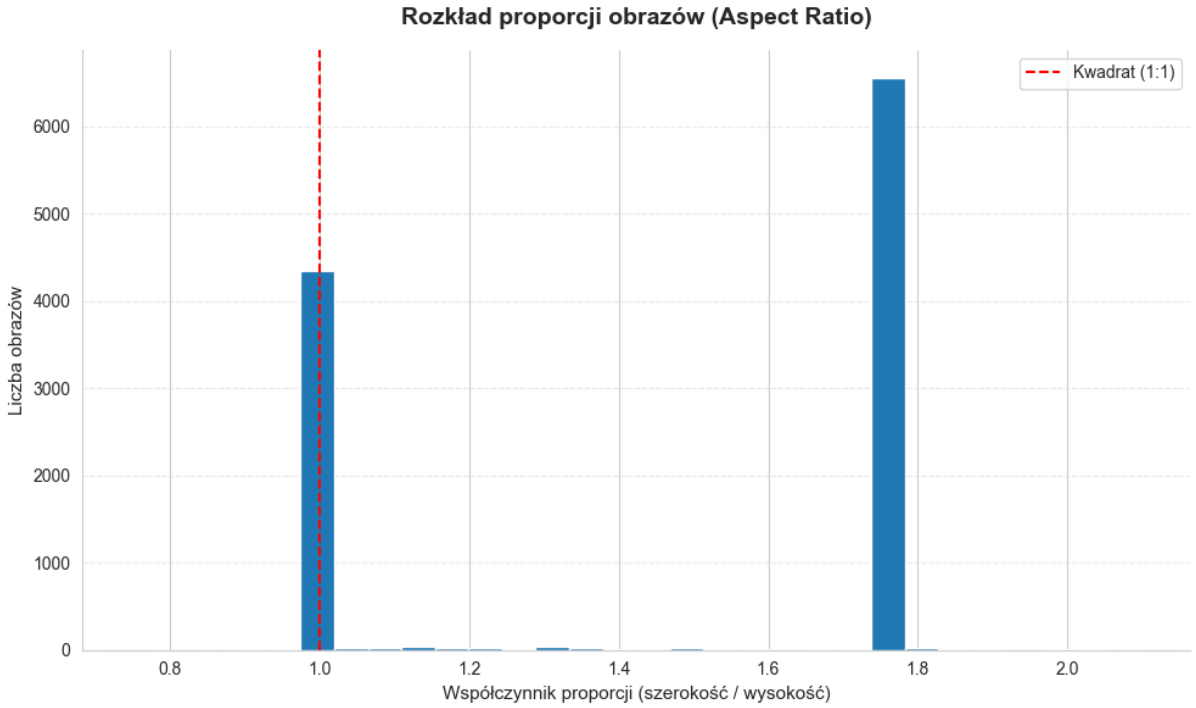
Szerokość (w px)	Wysokość (w px)	Liczność
1920	1080	5504
640	640	3997
1280	720	499
256	256	264
3840	2160	167

W strukturze zbioru wyraźnie dominują dwa standardy rozdzielczości: **Full HD (1920×1080)**, który obejmuje blisko połowę wszystkich próbek (~49%), oraz format kwadratowy **640×640**, stanowiący ponad jedną trzecią (~36%). Łącznie te dwie konfiguracje odpowiadają za blisko 85% całego zbioru.

Współczynnik proporcji (Aspect Ratio)

Widoczne są dwie dominujące grupy:

- Pierwsza grupa znajduje się w okolicach wartości 1.0, co oznacza obrazy zbliżone do **formatu kwadratowego 1:1**.
- Druga, znacznie liczniejsze skupisko występuje w okolicach wartości około 1.7–1.8, co odpowiada **formatowi 16:9 lub zbliżonym**.

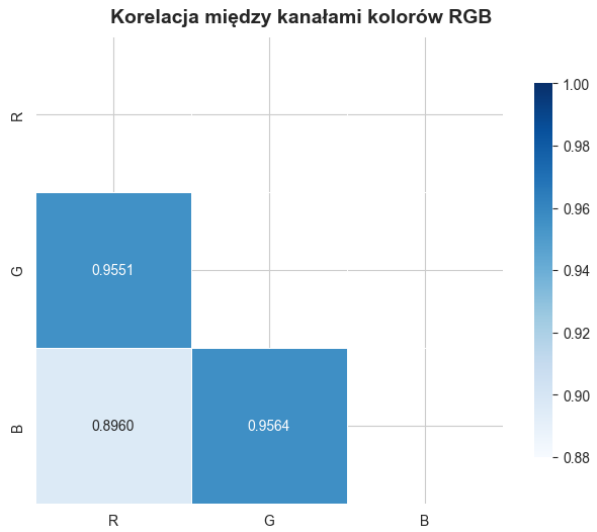


Kanały kolorów

Wyniki analizy wykazały, że wszystkie obrazy w zbiorze posiadają trzy kanały kolorów (RGB). Nie zidentyfikowano obrazów w skali szarości (1 kanał), formacie RGBA (4 kanały) ani w innych niestandardowych formatach.

Format	Liczność
GRAYSCALE	0
RGB	11249
RGBA	0
INNE	0

Pomimo że wszystkie obrazy zostały zapisane w formacie RGB, analiza różnic międzykanałowych wykazała, że **19,17% zbioru stanowią próbki o charakterze achromatycznym** (wizualnie czarno-białym), potwierdza to wysoka korelacja między poszczególnymi kanałami.

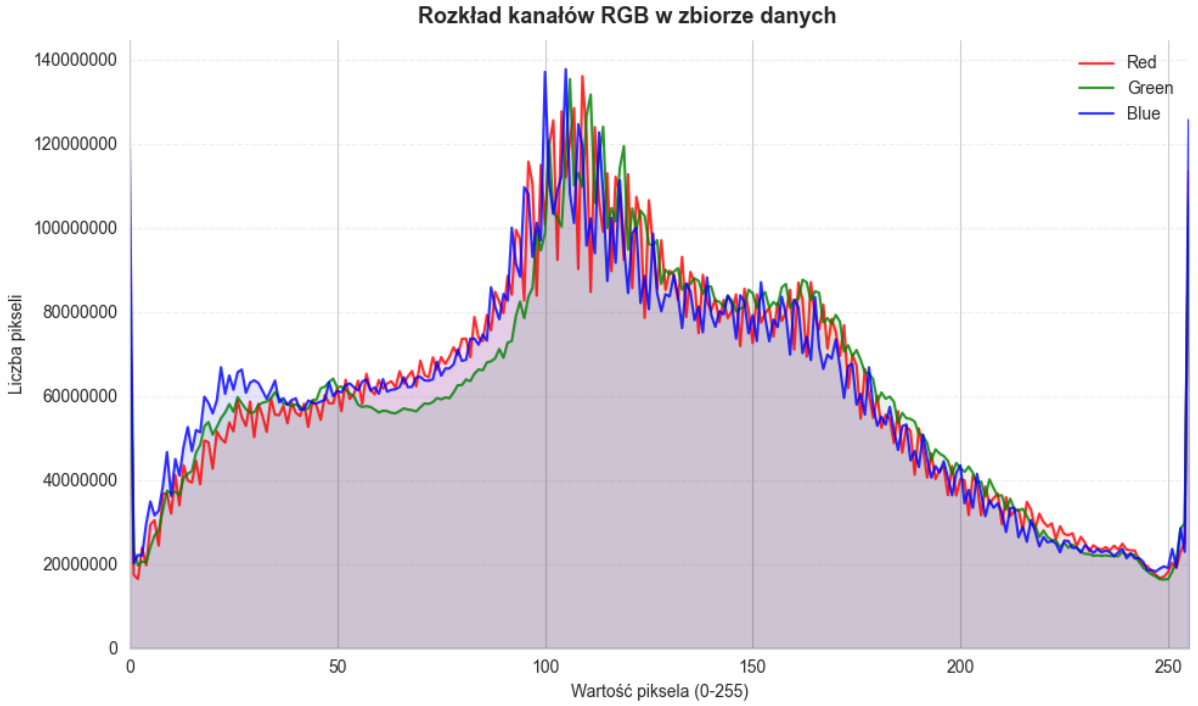


	R	G	B
ŚREDNIA	117.982	118.985	115.382
ODCHYLENIE STANDARDOWE	60.405	60.703	61.649

Wskaźniki statystyczne kanałów RGB

- **Średnia:** Wartości są do siebie zbliżone, oscyluja wokół połowy zbioru z lekkim przesunięciem w strone ciemniejszych tonów. Oznacza to że zbiór nie jest ani zbyt prześwietlony ani bardzo ciemny, kolory są zbalansowane.
- **Std:** Wartość ~61 sugeruje, że obrazy w zbiorze cechują się dużą zmiennością jasności.

Z wykresu przedstawiającego histogram wartości pikseli dla trzech kanałów kolorów (R, G, B), widzimy, że krzywe się niemal pokrywają, oznacza to, że w zbiorze niema dominacji jednego koloru, kolorystyka jest zrównoważona.



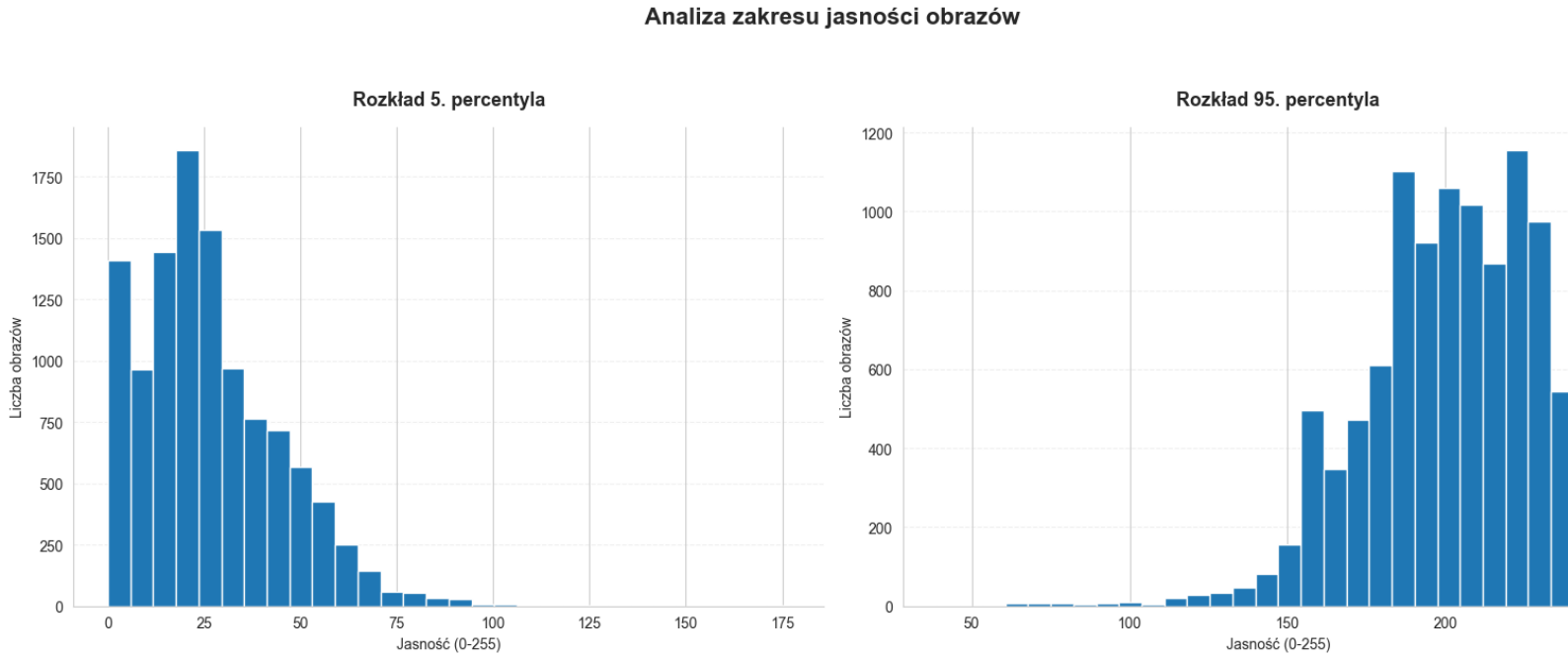
Jasność

W analizie jasności obrazów wykorzystano percentyle 5. i 95., które opisują odpowiednio poziom najciemniejszych oraz najjaśniejszych fragmentów obrazu.

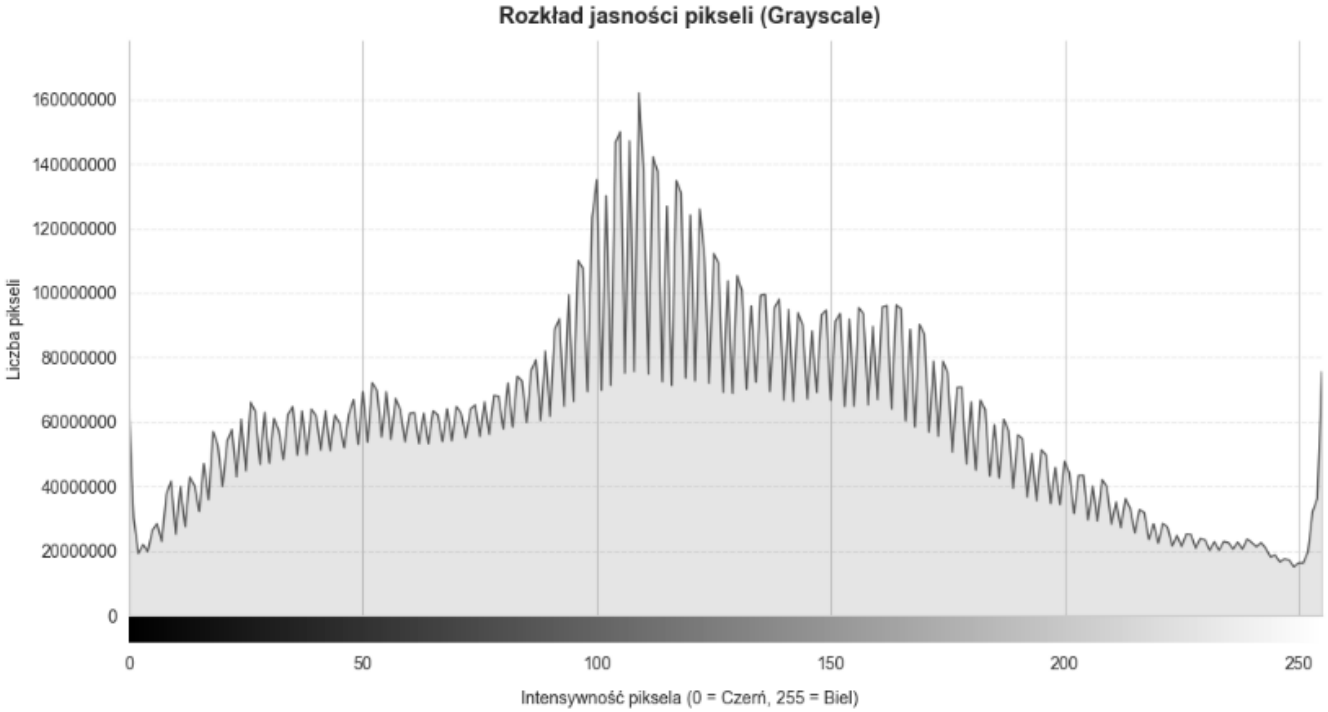
Wskaźniki statystyczne percentyli:

- **Średni 5. percentyl:** ~26,3
- **Średni 95. percentyl:** ~204,56
- **Szerokość tonalna (średni 95. percentyl - średni 5. percentyl):** ~178,3

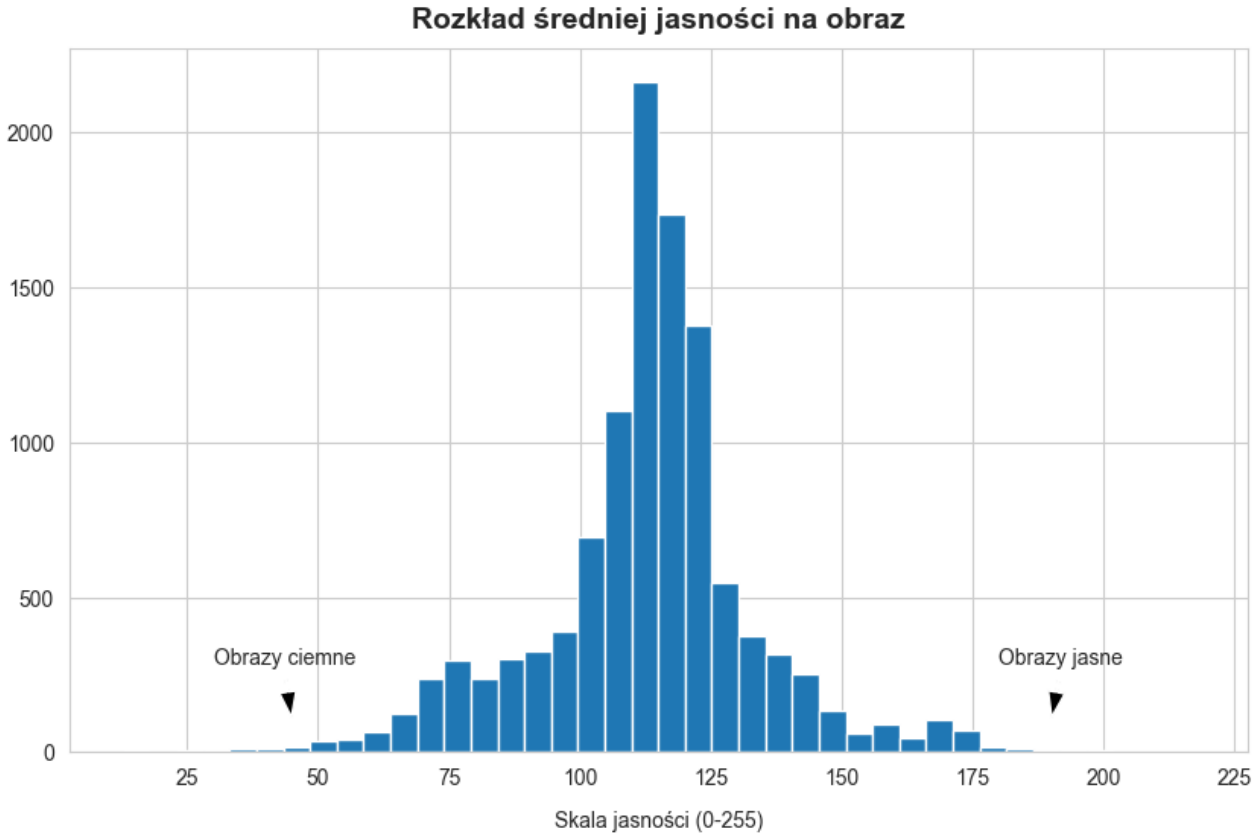
Z obliczonych wskaźników oraz wykresów rozkładu dla percentyli można wnioskować, że obrazy zachowują informacje (wysoką czytelność detali) zarówno w ciemnych jak i jasnych fragmentach. Wysoka szerokość tonalna sugeruje, że na zdjęciach wyraźny jest kontrast między ciemnymi i jasnymi partiami, co ułatwi modelowi poprawną separacje sylwetek od otoczenia.



Szeroki zakres rozkładu jasności pikseli pokazuje obecność zróżnicowanych warunków oświetleniowych (sceny dzienne oraz nocne).



Rozkład średniej jasności obrazów wskazuje, że większość próbek charakteryzuje się zrównoważonym poziomem luminancji, oscylującym wokół środkowych wartości. Ekstremalne przypadki - obrazy bardzo ciemne lub bardzo jasne – występują rzadko i stanowią niewielką cześć zbioru.

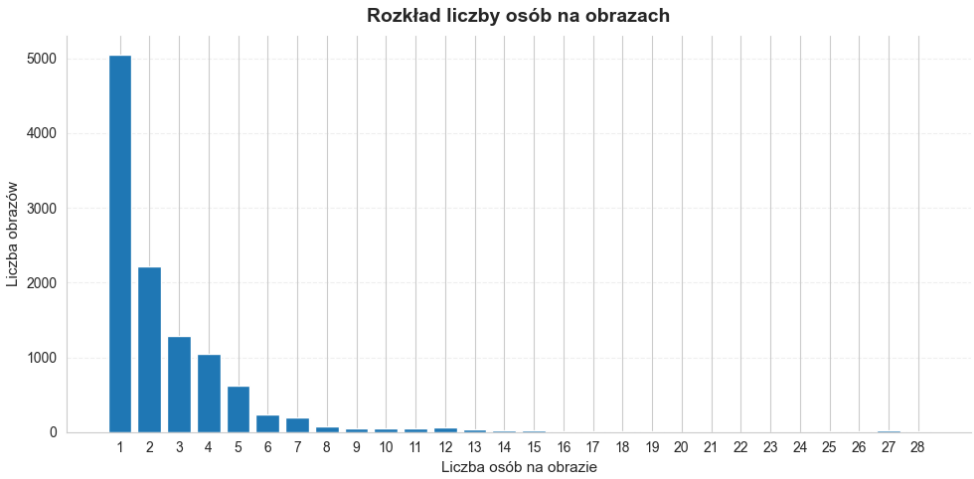


Dodatkowo przeprowadzona została analizę przypadków odstających pod względem jasności oraz kontrastu. Za obrazy bardzo ciemne uznano te, dla których 95. percentyl jasności był niższy niż 50, natomiast jako obrazy o niskim kontraście zdefiniowano próbki o szerokości tonalnej mniejszej niż 40.

W całym zbiorze zidentyfikowano jedynie 1 obraz bardzo ciemny oraz 3 obrazy o niskim kontraście. Oznacza to, że wizualnie nieczytelne lub problematyczne przypadki praktycznie nie występują, a jakość luminancji danych jest wysoka.

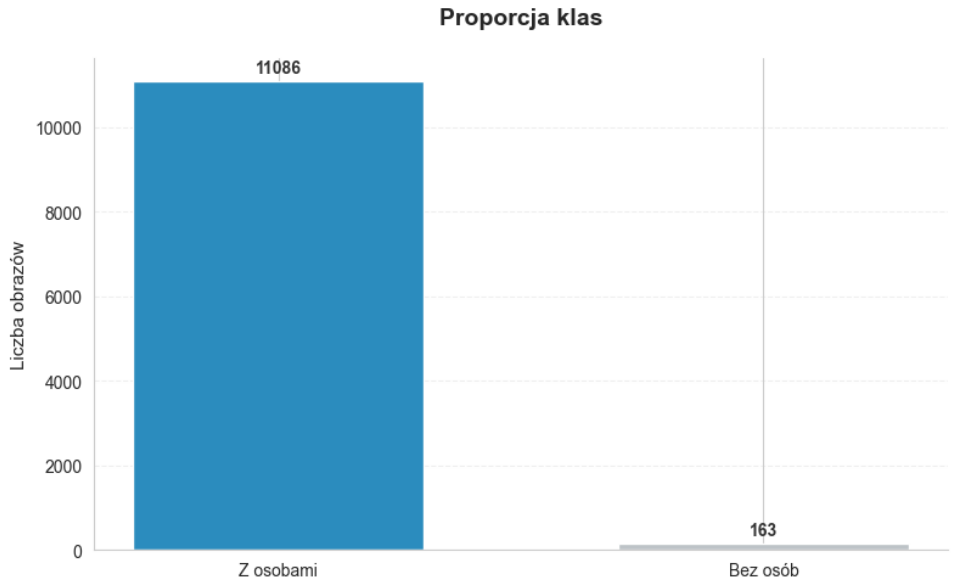
Etykiety (bounding box'y)

Wykres przedstawia liczbę obrazów w zależności od liczby oznaczonych osób (bounding boxów) w danej próbce. Widać, że dane głównie zawierają sceny o niskim zagęszczeniu, obrazy o wysokim zagęszczeniu stanowią mniejszość.



Zidentyfikowano **163** obrazy bez adnotacji: ~**1,5%** całego zbioru.

Podczas analizy wykryto **4** anomalie – zdjęcia bez etykiet pomimo widocznej obecności osoby na zdjęciu.



5. Interpretacja wyników

Rozmiary obrazów & Aspect ratio:

Analiza wymiarów obrazów wykazała, że zbiór nie jest jednorodny pod względem rozdzielczości ani proporcji. W trakcie budowania modelu trzeba pamiętać o ujednoliceniu danych wejściowych do tego samego rozmiaru.

Kanały kolorów:

Analiza wykazała silną korelację między kanałami kolorów, co wskazuje, że w wielu przypadkach informacje niesione przez kanały R, G i B są do siebie bardzo zbliżone. Znaczna część zbioru (ok. 19%) stanowią obrazy bliskie skali szarości. W kontekście detekcji obiektów nie stanowi to istotnego ograniczenia, ponieważ modele detekcyjne (np. YOLO) w dużej mierze opierają się na cechach strukturalnych, takich jak krawędzie i kształty, a informacja kolorystyczna ma zwykle charakter pomocniczy.

Jasność:

Analiza wykazała, że większość obrazów charakteryzuje się zrównoważoną jasnością oraz szerokim zakresem kontrastu. Rozkład percentyli wskazuje, że obrazy zachowują szczegóły zarówno w cieniach, jak i w jasnych partiach. Zbiór cechuje się dobrą jakością luminancji i nie wykazuje istotnych problemów związanych z niedoświetleniem ani prześwietleniem.

Etykiety:

Analiza adnotacji wykazała, że zbiór jest zdominowany przez obrazy zawierające niewielką liczbę osób. Oznacza to, że model będzie trenowany głównie na scenach o niskiej gęstości, co może ograniczać jego skuteczność w przypadkach silnie zatłoczonych scen.

Zidentyfikowano 163 obrazy bez osób. Jest to znikoma liczba w kontekście całego zbioru, w dodatku większość z nich nie odzwierciedla typowych warunków kamer CCTV, dlatego w dalszych etapach warto rozważyć ich usunięcie lub zastąpienie bardziej reprezentatywnymi przykładami tła.

Dodatkowo wykryto 4 przypadki błędnych oznaczeń - obrazy zawierające osoby bez przypisanych bounding boxów. Choć ich liczba jest marginalna, wymagają one korekty w celu zapewnienia pełnej spójności zbioru treningowego.

6. Odpowiedzi na pytania

Jaka jest jakość danych?

Jakość danych można ocenić jako wysoką.

Jakie wzorce i trendy występują w danych?

- Dominacja scen z małą liczbą osób (1–4).
- Dwie główne proporcje obrazów (format zbliżony do 1:1 oraz ~16:9).
- Silna korelacja między kanałami RGB.
- Około 19% obrazów zbliżonych do skali szarości.

Jakie problemy należy rozwiązać przed modelowaniem?

- Ujednolicenie rozmiaru wejściowego obrazów przy zachowaniu proporcji.
- Usunięcie błędnych adnotacji.
- Podjęcie decyzji dotyczącej obrazów bez osób (usunięcie lub zastąpienie bardziej reprezentatywnymi przykładami tła).

Jakie założenia okazały się prawdziwe, a jakie nie?

- Błędne założenie, że każde zdjęcie zawiera osobę.
- Poprawne założenie, że zbiór obejmuje różne warunki oświetleniowe.

Jakie nowe pytania pojawiły się podczas EDA?

- Czy model będzie skuteczny w scenach o dużym zagęszczeniu?
- Czy warto rozszerzyć zbiór o sceny bezosobowe?

7. Recomendacje

Strategie preprocessingu:

- Przy skalowaniu obrazów do stałego rozmiaru warto rozważyć technikę zachowującą proporcję unikającą deformacji sylwetek. Jedną z takich technik jest skalowanie typu **letterbox**.

- ➔ Z uwagi na niską rolę koloru w zbiorze warto sprobować technik augmentacyjnych, zmieniających kanały kolorów, np. **RandomGrayscale** polegającej na losowym przekształcaniu kolorowych obrazów wejściowych (RGB) w obrazy czarno-białe (w skali szarości) podczas procesu trenowania modelu. Random Grayscale uczy model, że błędy kolorystyczne nie mają znaczenia dla poprawnego rozpoznania człowieka.

Wybór modelu:

Dla problemów detekcji domyślnym wyborem są modele z rodziny YOLO.

Rozszerzenie zbioru:

W celu zwiększenia skuteczności modelu oraz poprawy jego zdolności generalizacji warto rozważyć:

- ➔ Zwiększenie liczby obrazów bez osób, co może przyczynić się do redukcji fałszywych detekcji.
- ➔ Rozszerzenie zbioru o sceny o dużym zagęszczeniu osób, aby poprawić skuteczność detekcji w bardziej złożonych przypadkach.