Paleta kolorów

Bartłomiej Szałach

25 listopada 2017

1 Wstęp

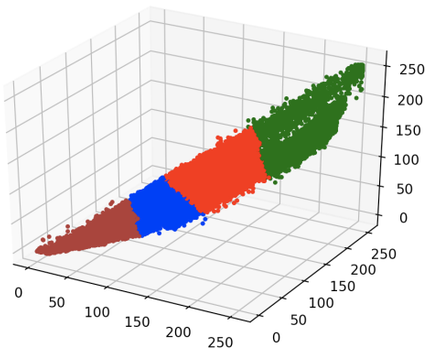
Celem zadania jest zbadanie palety kolorów występujących na obrazku. Do tego celu użyty zostanie *k-median clustering* oraz PCA. W zależności od doboru parametrów, takich jak obrazek, ilość klastrów lub usuwanie zduplikowanych kolorów postaram się porównać wyniki w różnych przypadkach.

2 Wyniki

2.1 Wstępne podejście

Rysunek 1: Ptak Kiwi 256x256

Dla k = 4 oraz wyboru losowych punktów jako środków otrzymano klasteryzację widoczną na Rysunku 2. Ze względu na "dużą" rozdzielczość obrazka i spowodowany przez nią długi czas działania programu obrazek do dalszych podejść zostanie zmieniony na inny.

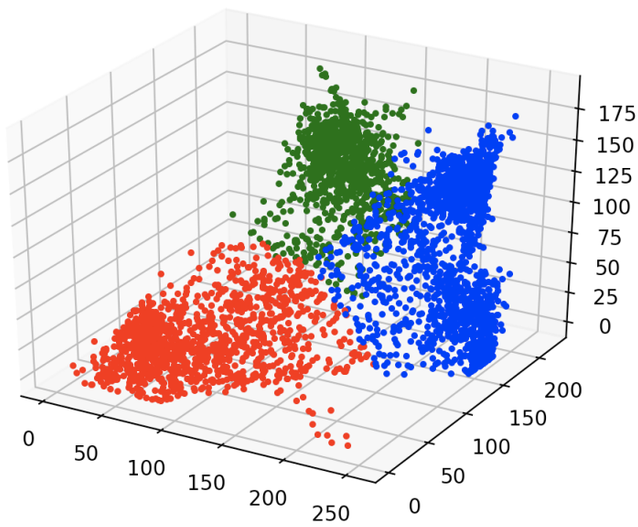


Rysunek 2: Klastry k-medians, k = 4

2.2 PCA + k-medians

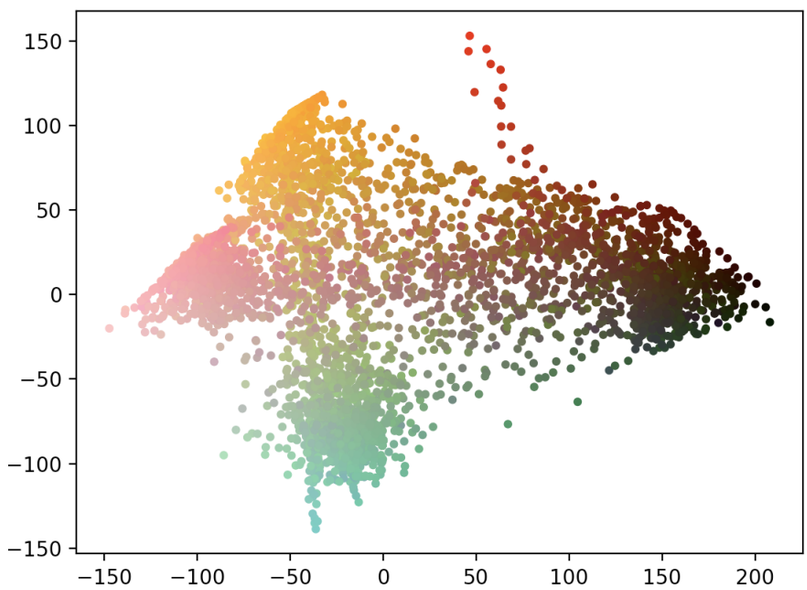
Rysunek 3: Marylin Monroe 64x64

Zdjęcie zostało zamienione na portret Marylin Monroe w niskiej roździelczości. Na obazku wyłaniają się 3 kolory: seledynowo-zielony w tle oraz na powiekach, ciemnożółty na włosach, oraz różowo-czerwony na twarzy i ustach. W oparciu o tą wiedzę dobrano parametr k = 3.



Rysunek 4: Klastry k-medians, k = 3

Na rysunku 4. pojawiają się obliczone klastry dla tego zdjęcia. Aby zwizualizować zastosowano PCA do danych wejściowych (Rysunek 5).

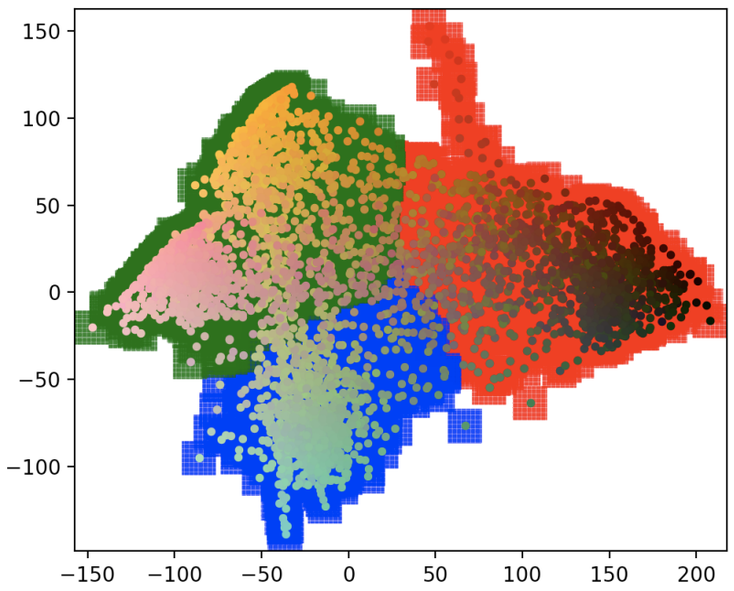


Rysunek 5: Dane po PCA do dwóch wymiarów

Kolejnym pomysłem było podbicie wymiaru punktów o grupę do której zostały przydzielone. I tak, 4096 3-wymiarowych punktów (r, g, b) zmapowałem ma punkty 4-wymiarowe (r, g, b, cluster). Na takich danych uruchomiłem jeszcze raz PCA, jednak otrzymany wynik był identyczny jak na Rysunku 5.

2.3 Podział na klastry i dobór parametru k

Na Rysunku 6 przedstawiona jest poprzednia już paleta kolorów, jednak tym razem wraz z klasyfikacją punktów do klastrów.

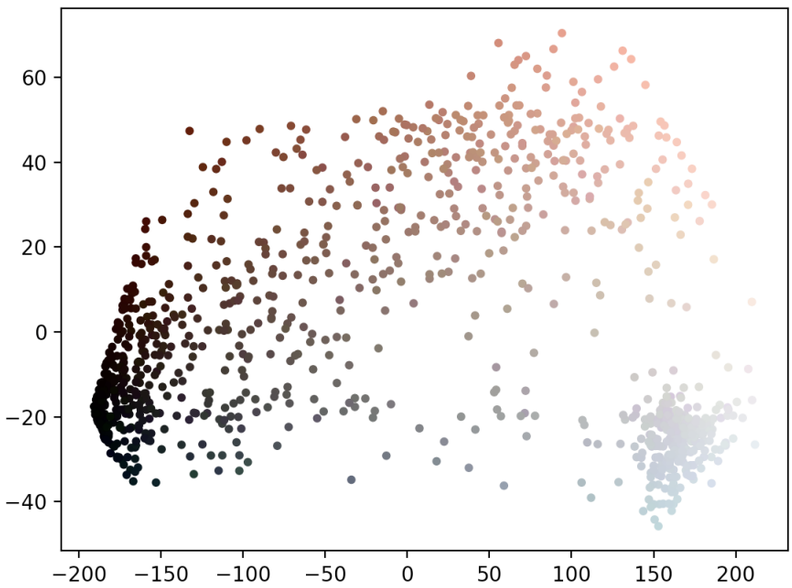


Rysunek 6: Klasteryzacja danych

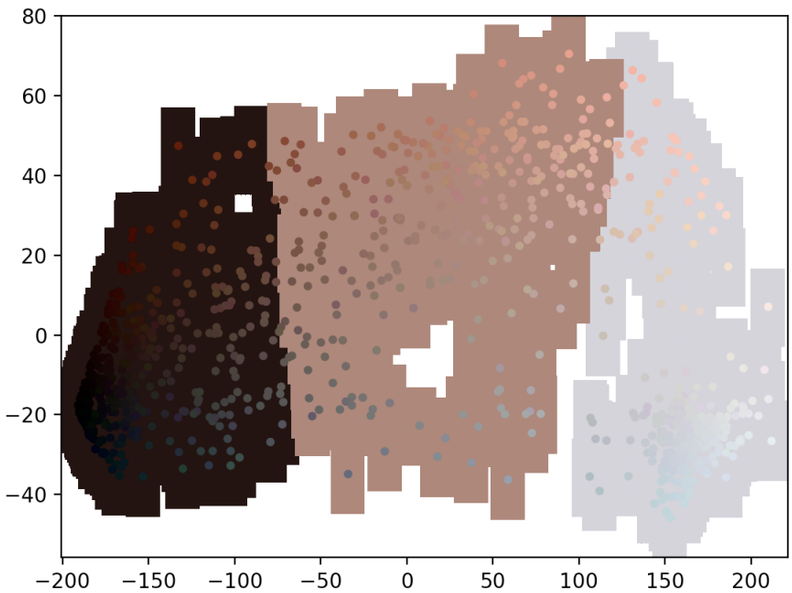
Niestety, ze względu na zbyt duże rozmiary naszej damy uruchomienie programu trwało około 10 - 15 minut na procesorze i5 7360U. Obrazek zostanie więc powtórnie zamieniony - tym razem na Hana Solo.

Rysunek 7: Han Solo, 32 x 32

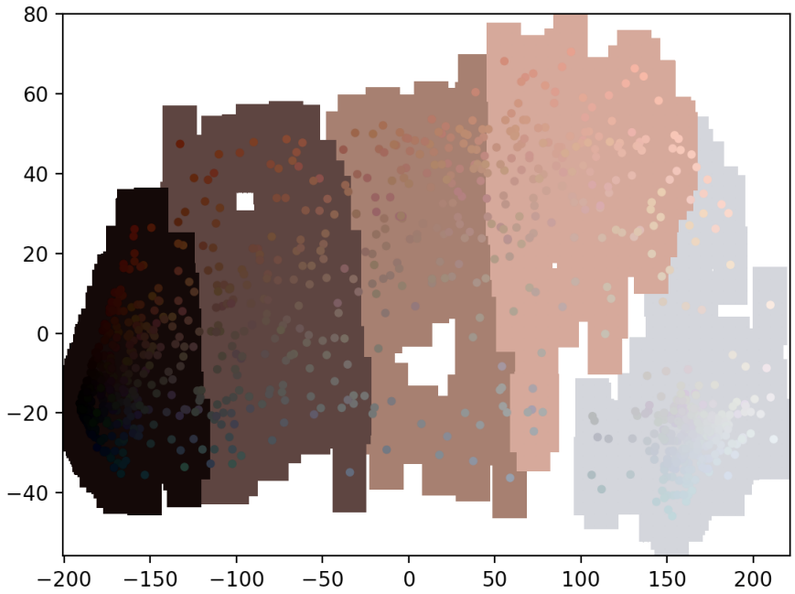
Analiza głównych składowych na pikselach Hana Solo daje rezultaty widoczne na Rysunku 8.
Kolejne obrazki będą przedstawiały klasteryzację dla różnych wartości parametru k. Kolor klastra jest medianą kolorów zaklasyfikowanych do danej grupy.



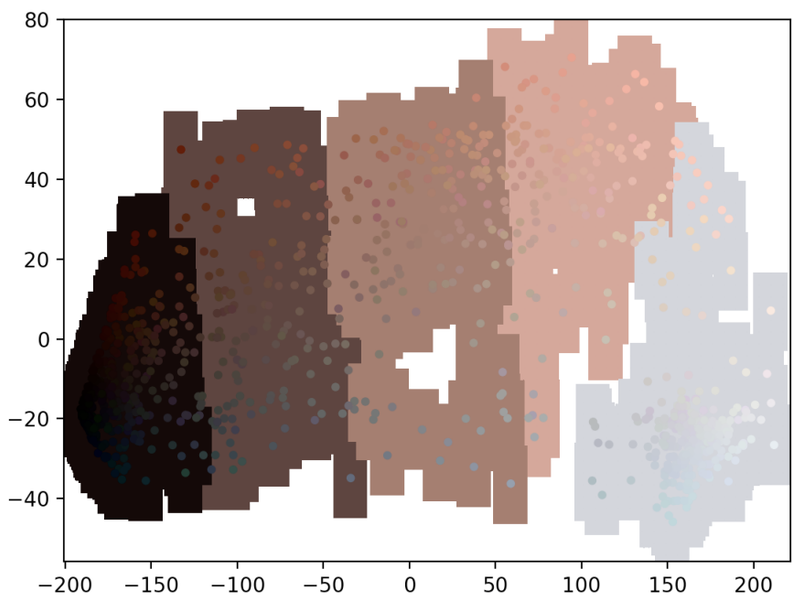
Rysunek 8: PCA



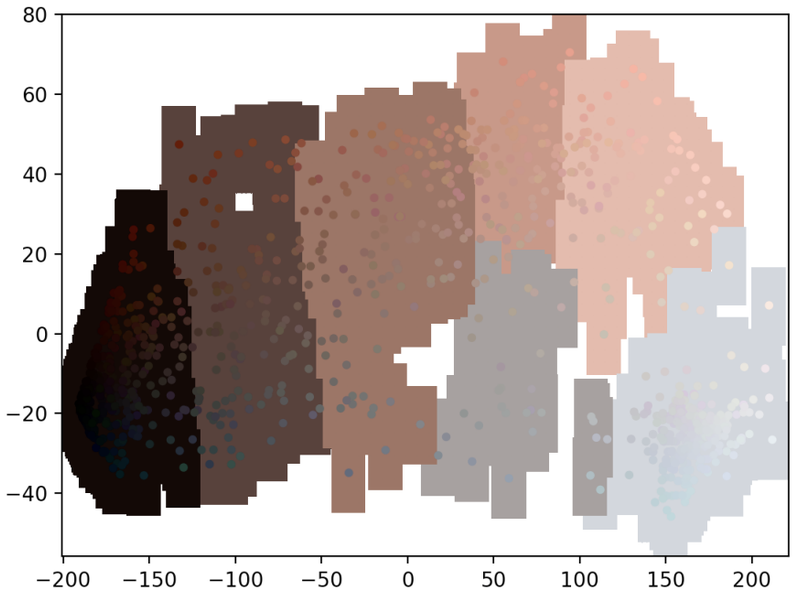
Rysunek 9: Klasteryzacja k = 3



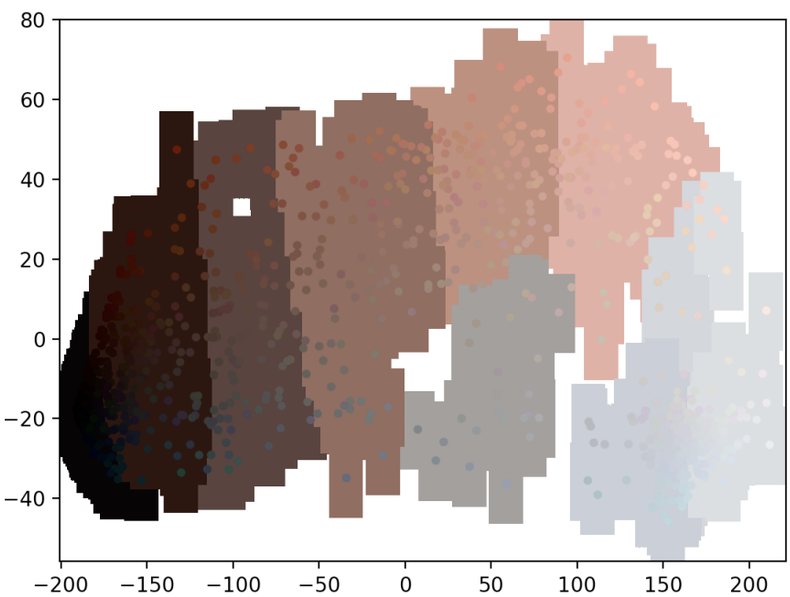
Rysunek 10: Klasteryzacja k = 5



Rysunek 11: Klasteryzacja k = 5 (druga wersja)



Rysunek 12: Klasteryzacja k = 7



Rysunek 13: Klasteryzacja k = 10

3 Podsumowanie i wnioski

- Dobór parametru k ma kluczowe znaczenie, gdyż określa on liczbę grup na które zostaną podzielone dane. Jednymi ze sposobów na dobór "dobrej" wartości jest wiedza domenowa o danych (na przykładzie Monroe) lub metoda empiryczna i porównywanie klasteryzacji dla różnych wartości (na przykładzie Solo).

- Wraz ze wzrostem parametru k, grupy stają się coraz mniej rozróżnialne, a obliczenia trwają dłużej. Zaobserwować można to dla k = 10.

- Dzięki zastosowaniu k-medians nie trafiamy w nieistniejące kolory, co pozytywnie wpływa na wizualizację palety kolorów.

- Jako środki zostały przyjęte losowe punkty ze zbioru. W związku z tym przy kolejnych uruchomieniach otrzymujemy różniące się klastry. Różnice są jednak małe, co pokazuje że algorytm jest w dużym stopniu odporny na losowy dobór, który w tych przypadkach wypadł dobrze.

- Przy tak małym zbiorze jak obrazek Hana Solo (1024 piksele) usuwanie duplikatów nie przynosi żadnych widocznych gołym okiem efektów. Po usunięciu takich samych pikseli ilość została zredukowana zaledwie do 987 pikseli. Z tego powodu klasteryzacja taka nie została nawet pokazana w sprawozdaniu, będąc uznaną przez autora za nieciekawą (aczkolwiek mogłaby być bardzo ciekawa dla większych zbiorów, w szczególności dla zdjęć modeli figur przestrzennych lub gradientów, a niekoniecznie dla twarzy)