Adrian Furman
Olga Kubiszyn
Gabriel Naleźnik
Konrad Walas
Jan Zajda
Wojciech Gomułka

Projekt zaliczeniowy

Analiza i Przetwarzanie Obrazów

Detekcja lokalizacji na podstawie analizy sekwencji wideo z kamer samochodowych

1. Problematyka projektu

Rozpatrywany problem pozwala na zastosowanie szerokiego wachlarza technik i metod znanych w analizie obrazów, a obejmujących przede wszystkim obszar rozpoznawania wzorców tekstu (*optical character recognition*). Zastosowanie znajdują również modele związane z klasyfikacją obiektów takich jak samochody (a przede wszystkim tablice rejestracyjne).

Cel projektu stanowiła estymacja lokalizacji nagrania, co potencjalnie mogło narzucać ograniczenia czasowe na zastosowane modele (program analizować musi wiele klatek dla każdego podanego na wejściu nagrania).

Trudność stanowi sam dobór kryteriów stosowanych w estymacji lokalizacji - najpierw należy wybrać wąski zakres rozważanych kryteriów przyporządkowania spośród szerokiego, trudnego do oszacowania zbioru. Przyporządkowanie może bowiem odbywać się na podstawie wykrytego na obrazie języka, kierunku jazdy pojazdów, tablic rejestracyjnych spotykanych w poszczególnych krajach, a nawet detali takich jak słupki i znaki drogowe, kolory linii na jezdni, spotykana roślinność i ukształtowanie terenu itp.

2. Kryteria przyporządkowywania lokalizacji

“Know-how” dla kryteriów stosowanych do odnalezienia prawdopodobnych lokalizacji, wywiedzione jest z ogólnej wiedzy geograficznej i technik, używanych m.in. przez graczy *Geoguessera* (gra internetowa polegająca na odgadywaniu lokalizacji wybranej losowo na Google Street View).

Poniżej przedstawiono kilka kryteriów oraz ich wady i zalety i uzasadnienie wyboru do stosowania w projekcie

Nazwa kryterium	Zalety zastosowania	Wady zastosowania	Uwagi
Język wykryty na klatce filmu	Wykrycie języka nietypowego (spotykanego np. w jednym kraju) znacząco ogranicza podzbiór możliwości.	Wykrycie języków o zasięgu globalnym (hiszpański, angielski, francuski) może doprowadzić do niepoprawnego określenia lokalizacji. Języki te występują na całym świecie. Język angielski może w dodatku być niejednokrotnie wykrywany, nawet jeśli nie jest językiem urzędowym na danym terytorium.	Zdecydowano się podjąć próbę wdrożenia w projekcie
Cechy tablic rejestracyjnych	Znakomite przy rozróżnianiu “podobnych” krajów Europy Zachodniej (Belgia / Holandia).	Element wymaga dużej dokładności przy wykrywaniu. W dodatku charakterystyczne cechy tablic często związane są z kolorem, który może ulegać zniekształceniom przy tak niewielkich obiektach.	Zdecydowano się podjąć próbę wdrożenia w projekcie
Kierunek ruchu pojazdów	Rozróżnienie Stanów Zjednoczonych i Kanady od innych krajów anglosaskich.	Wykrycie ruchu prawostronnego nie pozwala na zawężenie zbioru w znacznym stopniu. Z kolei ruch lewostronny jest domeną państw związanych historycznie z Wielką Brytanią (Irlandia, Indie, Australia, Nowa Zelandia), co niejednokrotnie nie niesie wystarczającej informacji	Zdecydowano się podjąć próbę wdrożenia w projekcie

		ze względu na podobieństwa między tymi państwami.	
Flora, ukształtowanie terenu	Teren górzysty może stanowić pomocniczą sugestię.	Na świecie przeważają strefy umiarkowanego klimatu, który trudno różnicować (ludzie osiedlają się w głównie w klimatach łagodnych lub ciepłych).	Bardzo trudne do automatyzacji
Wybrane cechy infrastruktury drogowej: znaki, słupki, kolory pasów.	Znakomite uzupełnienie, a czasami główne kryterium klasyfikacji.	Niezwykle dużo przypadków i możliwości.	Trudne do automatyzacji, wymaga odpowiednich modeli i dużo hard-codowanych cech specyficznych dla kraju.
Marki samochodów	W niektórych krajach o znaczących tradycjach motoryzacyjnych (np. Francja/Włochy) występuje silna reprezentacja aut marek rodzimych.	W wielu innych państwach trudno jednak doszukać się znaczącej przewagi danych marek lub wynika ona z czynników zupełnie odmiennych od tradycji motoryzacyjnych.	Wymagany byłby specjalistyczny klasyfikator.
Architektura	Charakterystyczna architektura w krajach skandynawskich (drewniane domy), “Nowy” Bliski Wschód (wieżowce), Ameryka Południowa - ceglane budynki.	W Europie jednolite budownictwo.	Z kamery samochodowej często niewiele można stwierdzić, w kwestii architektury.

3. Ekstraktory

3.1. Tekst ze znaków drogowych i reklam

Uznano, że jednym z najlepszych sposobów do poznania lokalizacji będzie czytanie tekstów ze znaków, reklam, plakatów itp. Dane te często dają nam wiedzę o języku używanym w miejscu wykonania nagrania. Dodatkowo, na niektórych znakach mogą znaleźć się bardzo pomocne informacje na temat nazw poszczególnych miejscowości.

Ze względu na dużą wartość wykrytych informacji tekstowych, do wykrywania tekstu na obrazach użyto dwóch sposobów.

Pierwszym jest zastosowanie wytrenowanego modelu *frozen_east_text_detection*. Stworzona z użyciem tego modelu sieć, bardzo dobrze sprawdza się do wykrywania na obrazach tekstu, który nie jest wyraźnie obrócony. Największym jednak problemem jest przetworzenie wykrytego tekstu na odpowiedni łańcuch znaków, w szczególności gdy wykryte słowa nie są zapisane w alfabecie łacińskim.

Drugim sposobem jest wykorzystanie biblioteki *easyocr.* Biblioteka ta w prosty sposób pozwala na znajdowanie tekstu na obrazach, a wyniki jej działania uznaliśmy za zadowalające. Wadą tego rozwiązania jest ograniczenie do kilku wybranych języków.
W projekcie zdecydowaliśmy się na wykrywanie języków japońskiego, chińskiego, arabskiego, ukraińskiego i angielskiego, aby móc wykryć różne rodzaje pisma.

Następnie przy pomocy biblioteki *langdetect* wykrywany jest język w znalezionym tekście, z procentową oceną pewności.

Rys. 1: Przykładowe zdjęcie ze znakiem informacyjnym.

Rys. 2: Tekst wykryty na podstawie Rys. 1, przy pomocy modelu *frost_east_text_detection*.

Rys 3. Przykładowa reklama.

Rys 4. Przykładowe rozpoznanie tekstu i języka japońskiego.

3.1.1. Wykrywanie miast/obiektów geograficznych

Po przeprocesowaniu klatki nagrania przez ekstraktor tekstu jako wynik otrzymujemy między innymi tablicę wykrytych napisów. Po wstępnym jej przeczyszczeniu (odrzuceniu liczb czy podejrzanie krótkich ciągów) wykonywana jest seria zapytań do API udostępniającego funkcjonalność geokodowania, czyli ustalenia współrzędnych geograficznych na podstawie danych takich jak nazwa miejsca, ulica czy kod pocztowy. W naszym przypadku będzie to zazwyczaj nazwa miasta lub miejsca o ile uda się taką informację z obrazu wydobyć. W odpowiedzi uzyskiwana jest lista potencjalnie dopasowanych punktów z bazy danych OpenStreetMap. Jako dostawcę usługi geokodowania wykorzystano Graphhopper.

3.2. Tablice rejestracyjne

Równie dobrym sposobem do przewidywania lokalizacji na podstawie informacji zawartych na pojedynczej klatce wideo jest analiza tablic rejestracyjnych samochodów.

Kolory oraz kształty tablic występujących na obrazie pozwalają bardzo dobrze zawęzić obszar poszukiwania państwa w którym został zarejestrowany materiał wideo. Można wyróżnić takie cechy charakterystyczne jak niebieski pasek po lewej stronie rejestracji europejskich, czerwone cyfry w Belgii, czy też całe żółte rejestracje w Izraelu oraz Wielkiej Brytanii.

Rys 5. Przykład rejestracji europejskiej. Niebieski pasek po lewej stronie.

Implementacja modelu do rozpoznawania tablic składa się z trzech etapów:

a) Wykrywanie tablic na zdjęciu

Detekcja tablic została zrealizowana za pomocą klasyfikatora kaskadowego Haara, przy pomocy funkcji cv2.CascadeClassifier. Do tego zadania został wykorzystany predefiniowany model klasyfikujący z pliku XML: “haarcascade_russian_plate_number.xml”. Funkcja detectPlates wykonuje metodę detectMultiScale na zdjęciu w skali szarości a następnie funkcja getPlates zwraca fragmenty bazowej klatki, zawierające tablice rejestracyjne.

b) Ekstrakcja cech ze zdjęcia tablicy rejestracyjnej

Na tablicach rozpoznanych w poprzednim kroku przeprowadzona zostaje ekstrakcja charakterystycznych cech. Większość tablic rejestracyjnych zawiera wąski, kolorowy pasek po lewej, bądź prawej stronie pozwalający na odróżnienie od siebie europejskich krajów. Cechą wykrywaną na zdjęciu tablicy był główny kolor dominujący w trzech obszarach:

Lewym pasku tablicy (10% szerokości tablicy), prawym pasku tablicy (również 10% szerokości) oraz na całej tablicy. Wykrywanie głównego koloru zostało zrealizowane przy pomocy metody K-średnich, wyodrębniającej główne barwy występujące na fragmencie rejestracji, a następnie wybraniu najliczniejszej klasy.

Rys 6. Obszary klasyfikacji barw na tablicy.
1 - lewy pasek, 2 - prawy pasek, 3 - Cała tablica

Niestety wykrywanie barw nie jest idealnym sposobem klasyfikacji tablic, ponieważ dużą rolę może odgrywać oświetlenie. W przypadku wyraźnego cienia lub światła padającego na rejestracje metoda wykrywająca barwy przydziela niepoprawny kolor.

c) Klasyfikacja tablicy do grupy krajów

Po wykryciu barw dominujących w poszczególnych częściach tablicy, następuje klasyfikacja wykrytej tablicy do grupy krajów o danej charakterystyce budowy rejestracji. Barwa, wykryta w poprzednim kroku w formacie HSV, zostaje przekazana do funkcji hsv_to_color, gdzie następuje przypisanie jej do jednego z predefiniowanych kolorów. Następnie, na podstawie 3 wykrytych kolorów, w funkcji predict_country_from_plate zostaje zwrócona grupa państw w których występuje charakterystyczna budowa tablicy wykryta na zdjęciu. Na przykładzie tablicy z rysunku x, W obszarze pierwszym wykryta została barwa niebieska, w drugim żółta, a na całej tablicy biała. Takie tablice występują w Portugalii, zatem wyjściem klasyfikatora w przypadku podania klatki zawierającej samochód portugalski jest zbiór jednoelementowy, złożony z identyfikatora Portugalii - “pt”.

3.3. Kierunek ruchu

Nadzieje na ograniczenie puli wyników pokładano również w detekcji kierunku jazdy. Zastosowana technika (która zostanie omówiona niebawem) nie pozwoliła jednak na uzyskanie satysfakcjonujących wyników, zwłaszcza w “złożonym”, miejskim otoczeniu. Rozważano również kilka innych podejść, które zostaną opisane w tym akapicie.

W zastosowanym rozwiązaniu przyjęto założenie, że wykryta liczba linii drogowych po prawej lub lewej stronie od środka obrazu, pozwoli oszacować, po której stronie jezdni znajduje się pojazd, a co za tym idzie - wyznaczyć czy ruch jest prawo, czy lewostronny. Wykrywanie odbywało się przy zastosowaniu probabilistycznej metody Hough, z ograniczeniem obszaru wykrywania adekwatnym do kształtu jezdni (implementacja na podstawie:
https://medium.com/analytics-vidhya/road-lane-detection-using-opencv-3e2d2740a4d)
Podejrzewanym od początku mankamentem tego rozwiązania była podatność na zwracanie fałszywych wyników w przypadku jezdni z wieloma pasami ruchu w jednym kierunku.

Z tego powodu sprawdzano skuteczność takiego podejścia do detekcji, w prostszych przypadkach (jezdnia o dwóch pasach w przeciwnych kierunkach). Wtedy użyta technika radziła sobie znacznie lepiej.

Okazało się jednak, że nie udało się przed oddaniem projektu zawęzić sposobu wykrywania linii tak, aby wykrywane były faktycznie jedynie linie na jezdni. Z tego powodu pojawia się wiele fałszywych detekcji, nawet przy użyciu gotowych sposobów zawężających obszar wykrywania linii do kształtu zbliżonego do trapezu (krawędzie jezdni "zbliżają" się do siebie ze względu na efekty związane z perspektywą).

Jedyne rozsądne wyniki uzyskiwano w zdjęciach z obszarów wiejskich i dróg leśnych (dwa pasy w przeciwległych kierunkach, brak infrastruktury mogącej powodować detekcje linii po bokach, wyraźnie zarysowane linie, mały ruch - brak przesłonięcia linii przez inne pojazdy).

Rys. 7. Prosty przykład z częściowo udaną detekcją linii drogowych

Na zamieszczonym powyżej rysunku uda się zwrócić poprawny wynik (linie wykryte po lewej sugerują ruch prawostronny przy naszym uproszczonym założeniu). Należy mieć jednak na uwadze dużą zależność tego rozwiązania od przyjętych parametrów (wiele linii pominięto choć powinny być wykryte). Dlatego w przeciwieństwie do rozwiązań związanych z OCR, przyjęty wpływ wykrytego kierunku ruchu na ostateczny wynik działania detektora lokalizacji jest niewielki.

Zamieszczony przykład należy więc traktować jako *proof of concept* rozwiązania na którym można byłoby potencjalnie bardziej polegać po odpowiednich usprawnieniach.

Poniżej przedstawiono inne rozważane podejścia do rozwiązania problemu detekcji kierunku ruchu i powody rezygnacji z nich:

Potencjalne rozwiązanie	Powód rezygnacji
Stosowanie różnic pomiędzy klatkami, aby wykryć ruch pojazdów i w ten sposób określić, które z pojazdów zbliżają się do obserwatora i czy odbywa się to z prawej czy lewej strony.	Obserwator nie jest statyczny, sam przemieszcza się pojazdem. Dlatego może dochodzić do wykrywania pozornego zbliżania się obiektów statycznych i/lub zbliżania się pojazdów, które są wyprzedzane przez obserwatora.
Zastosowanie gotowego modelu, który pozwoliłby wykryć pojazdy zwrócone frontem i na tej podstawie określić, po której stronie obserwatora się znajdują.	Nie udało się znaleźć modelu, który wykrywałby takie cechy pojazdów. Jego znalezienie mogłoby pozwolić osiągnąć przyzwoite działanie, nadal istniało jednak ryzyko wykrywania pojazdów zaparkowanych jako “jadących” z naprzeciwka.

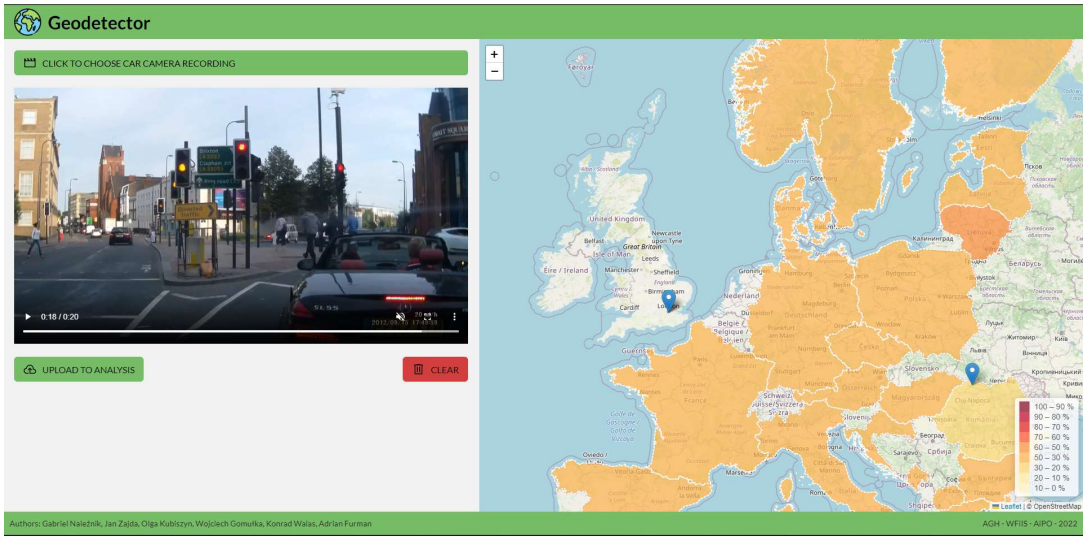
Nie znaleziono również modelu, który byłby w stanie rozwiązać ten konkretny problem (określenie kierunku jazdy - zbyt specyficzne zastosowanie).

4. Działanie aplikacji

4.1. Frontend

W ramach interfejsu użytkownika powstała prosta aplikacja webowa udostępniając formularz pozwalający na załadowanie wybranego nagrania za pomocą którego możemy nagranie przesłać do analizy na serwerze. Analizie poddana zostanie jedynie klatka obecnie widoczna w odtwarzaczu video więc istotne jest aby nagranie ustawić na takim momencie, który zawiera dobrze widoczne napisy, tablice rejestracyjne oraz pasy ruchu.

Po otrzymaniu odpowiedzi wynik jest prezentowany na mapie poprzez zaznaczenie krajów, z których według analizy może pochodzić nagranie. Im bliżej czerwonej bary jest kolor zaznaczenia kraju tym większe jest prawdopodobieństwo, że to właśnie z tego kraju pochodzi nagranie. Skala kolorów odpowiadajaca przedziałom prawdopodobieństwa znajduje się w prawym dolnym rogu mapy. Dodatkowo, jeśli z nagrania udało się wyodrębnić poprawne nazwy obiektów występujących w zasobach OpenStreetMap w ich miejscach przypinane są pinezki aby podjąć próbę dokładniejszego wskazania okolicy, z której może pochodzić nagranie. Po kliknięciu w pinezkę wyświetlany jest popup z nazwą dopasowanego miejsca.



Rys. 8. Interfejs aplikacji z widocznym rezultatem działania

4.2. Backend

Za część serwerową odpowiedzialny jest framework Flask. Całość jest bardzo prosta ponieważ wystawiany jest pojedynczy endpoint na który przesyłany jest wybrany film. Dodatkowo przesyłana jest informacja o wybranej chwili czasowej na fragmencie video oraz całkowitej długości filmu. W dalszym kroku klatka odpowiadająca wybranej przez użytkownika chwili czasowej poddawana jest działaniu kolejnych ekstraktorów. Wyniki ekstraktorów są zbierane w całość zgodnie z arbitralnie przyjętą formułą, która określa prawdopodobieństwo, że nagranie pochodzi z danego kraju:

$$0.7*L + 0.2*P + 0.1*R$$

gdzie ***L*** oznacza punkty przyznane za wykryty język, ***P*** za tablice rejestracyjne a ***R*** za kierunek jazdy.

Dla wykrytych na obrazie słów po wcześniejszym ich przefiltrowaniu pod kątem minimalnej długości oraz braku wystąpień liczb wykonywane są zapytania do serwisu geokodowania Graphhopper w celu uzyskania współrzędnych odpowiadającym odkrytym lokacjom. Po zgromadzeniu wszystkich danych zwracany jest rezultat w formie słownika krajów z przypisanymi im prawdopodobieństwami oraz listy pinezek, które zostaną wyświetlone na mapie:

Rys. 9.Struktura obiektu będącego rezultatem działania programu.

5. Podsumowanie

Rozpoznawanie geolokalizacji po wykonanym zdjęciu jest procesem bardzo złożonym i opartym na wielu klasyfikatorach cech szczególnych. Projekt opisany w niniejszym sprawozdaniu wykorzystał kilka z możliwych detektorów w celu najoptymalniejszego zawężenia kraju w którym zostało zrobione zdjęcie.

Początkowa implementacja polegała na przetwarzaniu wybranej ilości równomiernie rozłożonych na przesłanym nagraniu klatek, jednak ze względu na bardzo długi czas analizowania pojedynczej klatki oraz często występującą sytuację wybrania akurat takich klatek na których nia ma żadnych specyficznych cech mogących świadczyć o lokalizacji nagrania zrezygnowano z takiego podejścia. Ostatecznie analizie poddawana jest jedynie klatka obecnie widoczna na odtwarzaczu

Uzyskane wyniki pozostawiają wiele do życzenia i często wskazywany jest zupełnie błędny kraj. Dodatkowo następuje wyjątkowo duży rozrzut pinezek przedstawiających miejsca, których nazwy zostały odkryte na zdjęciu. Z dużym prawdopodobieństwem powodem takiego zachowania był wymóg podania modelom wykrywającym tekst języka, jakiego chcemy szukać na zdjęciu co prowadziło do wyszukiwania słów w danym języku niejako na siłę i skutkowało rozrzucaniem pinezek z wykrytymi miejscami po całej mapie.

Sposób implementacji zapewnia wygodną skalowalność aplikacji polegającą na możliwości dodawania kolejnych ekstraktorów cech, do których przekazywana byłby wybrana do analizy klatka.