Visión por Computador (2018-2019)

Grado en Ingeniería Informática y Matemáticas

Universidad de Granada

Descriptores HOG en la detección de peatones

Ignacio Aguilera Martos y Diego Asterio de Zaballa

15 de enero de 2019

Índice

1. Descripción del problema y enfoque de la resolución

En la resolución de este trabajo nos hemos planteado el siguiente problema a resolver, dada una imagen de una persona andando por la calle (un peatón), ¿cómo podemos reconocer que es un peatón?

Esta misma cuestión fue planteada por Navneet Dalal y Bill Triggs en su paper "Histogram of Oriented Gradients for Human Detection" en el que explican el desarrollo de unos descriptores que aplicados a su dataset de personas obtienen unos resultados muy buenos en la detección de las mismas.

Estos descriptores son los descriptores HOG o descriptores de histogramas basados en gradientes. Como veremos a lo largo del trabajo se han empleado distintas variaciones a la hora de hallar los descriptores por Dalal y Triggs, quedándonos nosotros con las elecciones que han resultado más fructíferas para ellos en su análisis.

Para comenzar hay que saber que la detección de personas es un problema difícil de abordar y que actualmente resulta muy interesante en aplicaciones por ejemplo en coches, de forma que si detecta un peatón andando por delante del vehículo este entienda que tiene que detenerse.

Las herramientas usadas en este proyecto han sido:

- Python para la implementación.
- OpenCV y NumPy para las operaciones de los algoritmos.
- El módulo de SVM incluído en OpenCV para poder predecir la existencia o no de un humano en una imagen en base a los descriptores calculados.
- El dataset dado por los investigadores empleado en la elaboración de su artículo.

A continuación hacemos una descripción un poco más elaborada del problema propuesto.

1.1. Problema a resolver

El problema consiste en, dada una imagen que contiene o no un peatón, debemos determinar si es que lo contiene y además una región aproximada en la que se encuentra el mismo.

Para la elaboración de la solución del problema hemos utilizado varios ingredientes entre los que tenemos la fase de creación de descriptores y entrenamiento de la SVM y la fase de test.

En la fase de entrenamiento de los descriptores obtenemos un dataset ya formado por parte del grupo de investigación con parches de personas y con parches que no son personas todos con el mismo tamaño. De esta forma podemos obtener descriptores comparables entre sí para las imágenes positivas y negativas con lo que entrenaremos una SVM.

La segunda fase es la de test que comparte dificultad e importancia con la primera. Como explicaremos más en detalle posteriormente resolvemos el problema de dada una imagen, ¿cómo la tenemos que analizar para poder identificar si hay o no personas y dónde están? En secciones siguientes veremos que la solución adoptada ha sido una ventana deslizante con algunas modificaciones importantes cuyos parches extraídos se pasan a la SVM y se predice sobre ellos para poder detectar el área de la imagen en la que obtenemos respuestas positivas a la pregunta de si hay o no una persona.

1.2. Descriptores HOG

El problema mencionado anteriormente se ha enfrentado desde distintos puntos de vista lo que ha dado lugar a diversos descriptores. Nuestra implementacion se basa en el paper de Navneet Dalal y Bill Triggs en el que desarrollan un descriptor basado en el gradiente de una imagen llamado descriptor HOG. El metodo se basa en calcular los gradientes de una imagen y a continuacion construir histogramas locales basado en la orientacion de dichos gradientes.

La idea basica sobre la que se desarrolla la tecnica es que las caracteristicas de forma y apariencia de un objeto de forma local se pueden resumir gracias a la distribucion de la orientacion de los gradientes de la imagen.

A continuacion se describe la implementacion de la tecnica a grandes rasgos. En primer lugar la ventana de la imagen sobre la que se va a construir el vector de caracteristicas se divide en celdas. Estas celdas son pequenas regiones de la imagen. En cada una de estas celdas se calcula el histograma de la direccion de los gradientes que en ella se encuentran. Esto define la re A continuacion estas celdas se agrupan en bloques para que sea posible normalizarlas.

diEl descriptor HOG es un vector de caracteristicas La construccion del descriptor HOG se apoya fundamentalmente en el gradiente de la imagen.

1.3. Fases de la resolución

1.3.1. Procedimiento para entrenar el modelo SVM

1.3.2. Procedimiento para la detección de un peatón en una imagen de test

El el paper no se discute en ningún momento el método de test utilizado por los investigadores por lo que, para que sea comparable con otras técnicas actuales, hemos empleado un método de test utilizado actualmente para medir el acierto de la detección.

En primer lugar se nos introduce una imagen de test de la que tenemos anotaciones en caso de ser positiva que nos indican el rectángulo de la imagen en el que tenemos un peatón. De esta forma una vez dada la región que hemos estimado para cada peatón podemos comprobar si se acerca a las regiones provistas por el fichero de anotaciones.

En primer lugar tomamos los tres primeros niveles de la pirámide Gaussiana para la imagen original, esto sería una lista con la imagen original y los dos niveles siguientes de la pirámide (incluimos a la imagen original como el primer nivel de la pirámide Gaussiana).

Para cada uno de estos niveles vamos a ir pasando una ventana deslizante de 128x64 píxeles y vamos a ir guardando estos parches. Tras la obtención de los parches calculamos sus descriptores y los pasamos a la SVM para que prediga si hay o no un peatón en cada parche. De esta forma finalmente obtendremos ventanas en las que sí pensamos hay un peatón y ventanas en las que pensamos que no hay un peatón.

Tomamos una matriz de ceros del mismo tamaño que el nivel de la pirámide Gaussiana que estamos evaluando y en cada posición sumamos 1 cada vez que obtengamos que dicho píxel está en una región en la que hemos predicho que hay un peatón. De esta forma terminamos con una matriz que tiene 0 en las posiciones en las que nunca se ha predicho que hay un peatón y números mayores que 0 indicando cuántas veces dicho píxel ha estado en una ventana con un peatón. Esta estructura es conocida como mapa de calor. Tras la obtención de esta matriz truncamos los valores menores que un umbral a 0 para quedarnos únicamente con regiones significativas, es decir, si tomásemos como umbral 2 entonces todos los píxeles con valor menor estricto que 2 tendrían un 0 en el mapa de calor.

Con esta estructura ya realizada podemos estudiar las regiones conexas de la misma, es decir, las regiones en las que tenemos números mayores que 0 adyacentes entre sí. De esta forma obtendremos regiones disjuntas del mapa de calor. Sólo resta buscar el rectángulo más pequeño que engloba dicha región y ya tendríamos nuestra predicción de la región de la imagen en la que tenemos peatones.

Por último se comprueba que las regiones encontradas solapen al menos en un 50 %

con las regiones aportadas por los ficheros de anotaciones y devolvemos para cada imagen $\frac{número\ de\ peatones\ acertados}{número\ de\ peatones\ totales}$. De esta forma tenemos una medida entre 0 y 1 que nos dice cuánto hemos acertado en cada imagen. Si sumamos todos estos números tendríamos una medida global del acierto, de forma que la puntuación máxima que podríamos obtener sería el número de imágenes.

De igual modo para hacer una predicción y medida equivalente en las imágenes negativas, para cada imagen de test negativa tomamos 10 ventanas de tamaño 128x64 de forma aleatoria y devolvemos el número de ventanas acertadas partido por el número de ventanas totales, en este caso 10. De esta forma obtenemos que las medidas de las imágenes negativas y positivas acertadas serían comparables entre sí.

Al realizar el análisis de los peatones en cada uno de los niveles de la pirámide Gaussiana obtenemos una ventaja, y es que los posibles peatones que haya en la misma serán de menor tamaño y por tanto será más probable que el peatón nos entre por completo en una ventana al ir avanzando en los niveles de la pirámide con lo que el parche que obtengamos será más parecido a los pasados para entrenamiento de la SVM.

Así mismo hemos realizado un test ortodoxo del modelo tomando imágenes ya recortadas del test de tamaño 128x64 con personas y sin personas de forma que es capaz de decirnos si hay o no personas en esa imagen (sólo hay una persona por recorte). Este test nos proporciona también una medida del método que subyace a lo realizado en el proceso anterior. Podríamos decir que este proceso de test de los recortes realizados por nosotros a mano empleando los ficheros de anotaciones es la parte del reconocimiento de los parches empleada en el método anterior al obtenerlos con la ventana deslizante. En la siguiente sección detallaremos más en profundidad los resultados obtenidos y los comentaremos.

2. Valoración de los resultados

3. Trabajo futuro: propuestas de mejora