Universidad de Granada

Ingeniería Informática

Computación y Sistemas Inteligentes

Cuestionario 3

Autor: José Antonio Ruiz Millán
Asignatura: Visión por Computador
21 de diciembre de 2018

1. **¿Cuáles son las propiedades esenciales que permiten que los modelos de recuperación de instancias de objetos de una gran base de datos a partir de descriptores sean útiles? Justificar la respuesta.**

   La propiedades se centran básicamente en el funcionamiento de la misma, podemos decir que las principales propiedades son **el uso de indices invertidos**, el uso de **palabras** que constituyen la **bolsa de palabras** y los propios **descriptores SIFT** asociados a cada una de las regiones **(parches)** de las imágenes.

   Con esto, tenemos una estructura que nos permite después de haber creado los diccionarios con la gran base de datos de la que partimos y tener todos los descriptores asociados a las regiones y los distintos parches, poder crear el indice invertido y la bolsa de palabras haciendo que estos tipos de modelos sean robustos para la recuperación de instancias de objetos.

2. **Justifique el uso del modelo de bolsa de palabras en el proceso de detección y reconocimiento de instancias de objetos ¿Qué ganamos?, ¿Qué perdemos? Justificar la respuesta**

   La **bolsa de palabras** es el conjunto de palabras que aparecen en una imagen concreta, siendo cada palabra una caracteristica/zona relevante de la imágen. Esta bolsa de palabras se puede representar como un histograma ya que no sólo contiene las palabras que aparecen en la imagen si no que también almacena la frecuencia de cada una de las palabras.

   Esto nos permite hacer un resumen de la propia imagen y así poder **comparar dos imágenes utilizando sus bolsas de palabras** y no sus descriptores directamente.

   Gracias a el uso de bolsas de palabras, tenemos algunas **ganancias** como pueden ser:

   - Es flexible a la geometría, a deformaciones y a cambios del punto de vista.
   - Nos permite hacer un resumen de la imagen bastante representativo.
   - Permite tener la inforamción de la imagen estructurada de forma vectorial (histograma).
   - Tiene buenos resultados en la práctica.

   Por otro lado, tenemos algunas propiedades que **no son tan buenas** como pueden ser:

   - Cuando la bolsa de palabras cubre toda la imagen, se produce una mezcla entre el fondo y el primer plano.
   - Crear un diccionario óptimo no es trivial, por lo que encontrar un diccionario óptimo es complicado.
   - El modelo básico ignora la geometría, se necesitan verificar posteriormente.

3. **Describa la diferencia esencial entre los problemas de reconocimiento de instancias y reconocimiento de categorías. ¿Qué deformaciones se presentan en uno y otro? Justificar la respuesta**

   La diferencia principal entre entos dos problemas es el propio **reconocimiento del objeto**, es decir, en los problema de reconocimiento de instancias estamos buscando una instancia/objeto concreto, con una forma una posición concreta, sin embargo, en los problemas de reconocimiento de categorías estamos buscando el objeto en sí como clase/categoría pero no un objeto concreto. Por ejemplo en el caso de una silla, en reconocimiento de instancias

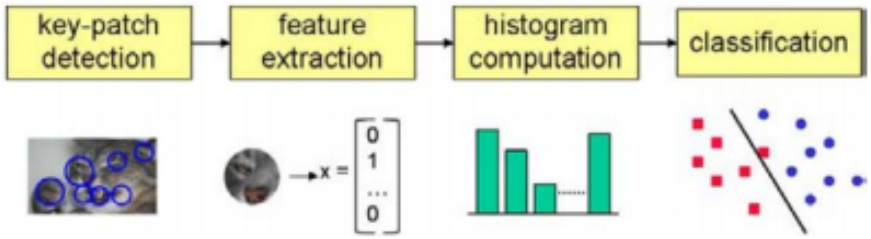
podemos buscar una silla concreta en una imagen, pero si lo que queremos es reconocer sillas dentro de una imagen, este problema es de reconocimiento de categorías ya que estamos intentando detectar una silla independientemente de su forma de su color y de su posición, lo que queremos es detectar cualquier tipo de silla. Esto hace que los problemas de reconocimiento de categorías sean problemas más complejos que los de reconocimiento de instancias.

Respecto a las **deformaciones**, en reconocimiento de instancias pueden parecer deformaciones en la escala o en la rotacion y cambios de iluminación. En reconocimiento de categorías aparecen además cambios en la forma geométrica ya que una silla puede tener 4 patas y otra 2.

4. **¿Es posible usar el modelo de bolsa de palabras para el reconocimiento de categorías de objetos? Justificar la contestación**

   Sí, este algorimto es uno de los algoritmos más simples para el reconocimiento de categorías y también es conocido como *bolsa de puntos clave (keypoints)* o *bolsa de características*.

   Este algoritmo simplemente calcula la distribución (histograma) de las palabras visuales encontradas en la imagen y compara esta distribución con las encontradas en las imágenes de entrenamiento.



5. **Suponga que desea detectar, en una imagen, una instancia de un objeto a partir de una foto del mismo tomada desde el mismo punto de vista del que aparece en la imagen y en un entorno de iluminación similar. Analice la situación en el contexto de las técnicas de reconocimiento de objetos e identifique que algoritmo concreto aplicaría que fuese útil para cualquier objeto. Argumente porqué funcionaría y especifique los detalles necesarios que permitan entender su funcionamiento.**

   Utilizaria **descriptores SIFT** para el emparejamiento. Tenemos una escena en la que el objeto que queremos buscar está en la misma posición, misma escala y con la iluminación similar que la imágen que tenemos sobre él. Podríamos utilizar tanto la bolsa de palabras como SIFT, pero he decidido elegir SIFT por su simplicidad y porque su funcionamiento para este problema concreto es más que satisfactorio.

   Este algoritmo **funcionaría** porque como he comentado anteriormente estamos en una situación en la que la imagen que tenemos sobre el objeto está tomada desde el mismo punto de vista por que mantiene el escalado y la rotacion intacta y estamos en dos imágenes con iluminación similar, por lo que éste algoritmo funcionaría perfectamente.

   El funcionamiento del mismo es el siguiente:

   - Si la imagen del objeto que tenemos es únicamente el objeto podemos pasar al siguiente paso, pero si no lo es, primero debemos crear una máscara sobre la imagen para

únicamente calcular los descriptores de esa zona de la imagen.

- Una vez tenemos eso, lo que hacemos es ejecutar el algoritmo SIFT sobre las dos imágenes para obtener sus keypoints y sus desciptores. Después ejecutamos un algoritmo de emparejamiento como puede ser *2NN-Lowe* para obtener las correspondencias entre las dos imágenes (desde la imagen del objeto hacia la imagen total).
- Una vez tenemos esto ya podemos saber a través de las correspondecias dónde se encuentra el objeto en la segunda imagen a través de los puntos de la primera imagen.

6. **Suponga de nuevo el problema del ejercicio anterior pero la foto que le dan está tomada con un punto de vista del objeto distinto respecto del objeto en la imagen. Analice que repercusiones introduce esta modificación en su solución anterior y que cambios debería de hacer para volver a tener un nuevo algoritmo exitoso. Justificar la respuesta.**

   En este caso el problema cambia radicalmente ya que ahora tenemos una deformación entre el objeto que nosotros tenemos y cómo está ese objeto en la imagen. Esto afecta a la solución del ejercicio anterior ya que ahora si utilizasemos los descriptores SIFT no tendríamos ninguna o muy pocas correspondencias.

   Por ello, para este caso no podemos utilizar simplemente los descriptores SIFT y pasamos a utilizar **bolsas de palabras** ya que ésto nos permite generalizar y utilizar las palabras para comprobar si el objeto está realmente en la segunda imagen. Este algoritmo lo podemos utilizar porque es flexible a estos tipos de cambios como ya indique en el ejercicio 2. Claramente para poder utilizar éste algorimo necesitamos tener la estructura del mismo como se comentaba en el ejercicio 1 ya que sin esa estructura no tenemos las palabras ni los descriptores asociados a cada una de las palabras.

   Una vez tenemos esto, utilizando las palabras que contiene la imagen del objeto y las palabras que tiene la imagen destino podríamos comprobar si ese objeto está en la imagen.

7. **Suponga que una empresa de Granada le pide implementar un modelo de recuperación de información de edificios históricos de la ciudad a partir de fotos de los mismos. Explique de forma breve y clara que enfoque le daría al problema. Que solución les propondría. Y como puede garantizar que la solución podrá ser usada de forma eficiente a través de dispositivos móviles.**

   Tenemos un problema de reconocimiento de instancias ya que queremos detectar si en una imágen se encuentra un edificio concreto. Para ello, tenemos un abanico de soluciones donde principalmente tenemos los desciptores SIFT, bolsa de palabras y redes neuronales. Sabemos que las imágenes van a tener deformaciones por lo que los descriptores SIFT están descartados. Por otra parte sabemos que necesitamos que se pueda ejecutar en un dispositivo móvil, eso implica cómputo y memoria para el dispositivo. Sabemos que las redes neuronales necesitan un cómputo muy elevado y una cantidad de imágenes para entrenamiento muy grande, por ello descarto esta solución y me quedo con **la bolsa de palabras**.

   Como he comentado, propondría como solución la **bolsa de palabras**, ya que para las imágenes de entrenamiento, al tener las bolsas de palabras las tenemos comprimidas y el cómputo que tenemos que realizar es mucho menor que para una red neuronal. No obstante sabemos que éste problema es un problema dificil de abordar por la cantidad de imágenes

que necesitamos para entrenar el modelo.

Una posible **mejora** es añadirle verificación espacial basada en el algoritmo *RANSAC* y así obtener una mejora cuando tengamos que decidir qué edificio estamos viendo y descartar edificios que estén alejados del que buscamos.

Podemos **garantizar** su funcionamiento en un dispositivo móvil como he comentado anteriormente porque el cómputo que necesita realizar es pequeño y puede ser ejecutado por los procesadores que tienen los disporivos móviles actuales. Tambien necesitamos tener memoria en el dispositivo, esto nos pone el problema de que a más imágenes de entrenamiento más imágenes necesitamos almacenar en el dispositivo y puede llegar a ser pesado, no obstante al tener las bolsas de palabras, tenemos como un resumen de las propias imágenes y no necesitamos todas las imágenes en sí.

8. **Suponga que desea detectar la presencia/ausencia de señales de tráfico en imágenes tomadas desde una cámara situada en la parte frontal de un coche que viaja por una carretera. Diga que aproximación usaría y porqué. Identifique las principales dificultades y diga como las resolvería. Los argumentos deben ser sólidos y con fundamento en las técnicas estudiadas.**

   Para este problema se me ocurren principalmente dos opciones, una **red neuronal convolucional**[6] y el algoritmo ***Modified Hough Transform*** (MHG)[5].

   En primer lugar **la red neuronal convolucional** puede ser entrenada con millones de imágenes sobre señales de tráfico de todo tipo e incluso con señales que tengan obtáculos o no se vean demasiado bien. Estos algoritmos gracias a su funcionamiento por capas y el *backpropagation* permiten darnos un *accuracy* bastante alto y en tiempo real cosa que es muy importante para este tipo de desarrollos. Con un entrenamiento bastante duro podríamos ajustar mucho la red hasta conseguir muy buenos resultados y únicamente meter la red en el coche para que una vez tenga la imágen, sea capaz de clasificar señal o no señal en tiempo real y ejecutar una determinada acción. Como he comentado la potencia de este método es que como ya lo tenemos entrenado no necesita un procesamiento en el coche y si ha sido bien entrenado, otra caracteristica potencial es que será capaz de diferenciar señales que no se visualicen bien aunque estén algo tapadas o que no se vea el total de la señal, siempre y cuando se le hayan proporcionado un buen conjunto de entrenamiento.

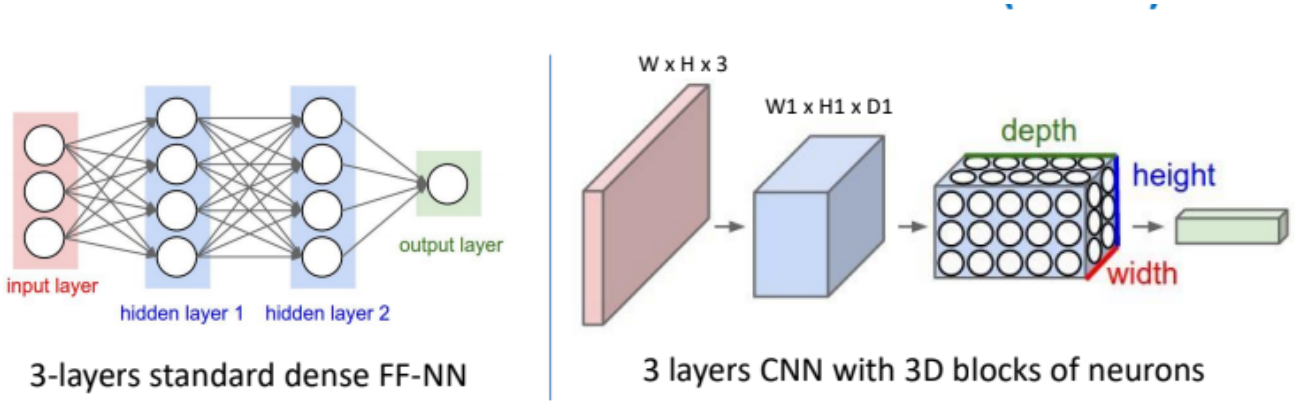
   El segundo caso es un caso en el que no hemos profundizado en clase, no obstante es un método utilizado como se puede ver en el paper. Este método consigue reconocimiento de imágenes en tiempo real he incluso en imágenes en las que la velocidad del coche haga que se vean un poco borrosas. Se utiliza un umbral de distancia para aceptar una señal como válida ya que cuando se encuentra muy lejos puede dar falsos positivos. Como el anterior, es un algoritmo que consigue darnos el resultado en tiempo real y permite la detección de señales con deformaciones aunque no es tan robusto al ruido como las redes neuronales convolucionales.

   Como podemos ver los dos métodos cumplen las caracteríticas principales para estos tipos de problemas, que son **la respuesta en tiempo real** y que sean capaces de **detectar señales parciales** o señales con ruido. No obstante como el enunciado nos advierte, estos algoritmos son algoritmos de aproximación y no son algoritmos exactos que están dá a día mejorando para llegar al punto más alto dentro de una aproximación.

9. **¿Qué han aportado los modelos CNN respecto de los modelos de reconocimiento de objetos empleados hasta 2012? Enumerar las propiedades comunes entre ellos y aquellas claramente distintas que hayan permitido una mejora en la solución del problema por parte de las CNN. Dar una opinión razonada de por qué significan realmente una mejora.**

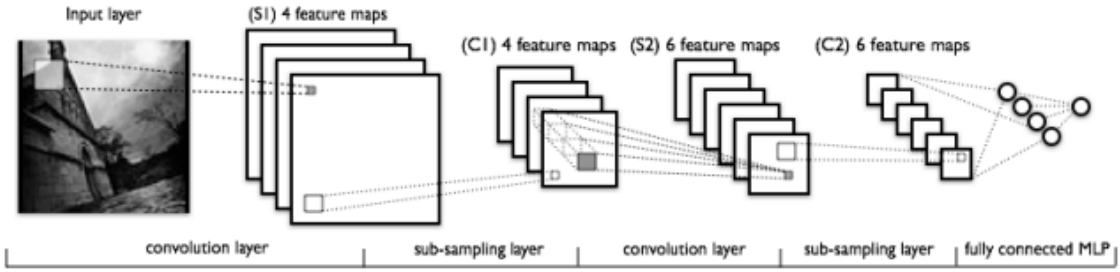
   La principal aportación de los modelos CNN es el **modelo de convoluciones** que se aplican antes de las últimas capas formadas por una *full connect*. Gracias a este descubrimiento, Google fué capaz de arrasar con todos sus competidores mejorando en un porcentaje bastante elevado la precisión de acierto con la base de datos de ImageNet. Otra aportación es la **eliminación de preprocesado de imágenes** respecto a otros modelos no basados en redes. Antes de esta aportación **las redes eran un conjunto** de nodos formados por capas donde todos ellos estaban interconectados entre sí y se mandaban información, usando *backpropagation* para difundir en error a la capas anteriores y mejorar el ajuste de la función. Estas son algunas de las características que **tienen en común** ya que aún tenemos en las últimas capas un modelo *full connect* y seguimos utilizando *backpropagation* para difundir el error hacia las capas anteriores.

   Sin embargo, lo que **las diferencia** es las capas que tienen antes de la *full connect* ya que anteriormente eran así todas las capas y a partir de ese momento consiguieron una gran mejora añadiendo un proceso de convolucion+downsampling hasta llegar a la últimas capas full-conect y clasificar la imánge. Esta característica fué un gran avance en las resdes neuronales para el reconocimiento en imágenes ya que anteriormente si la red tenía muchas capas o muchos nodos por capa, se incrementaba la dificultad de aproximar la función por la cantidad de conexiones que teníamos.



3-layers standard dense FF-NN

3 layers CNN with 3D blocks of neurons



Simple example of CNN

Como vemos en la imágen lo que conseguimos es resumir la imágen a un tamaño más pequeño donde la información más importante para clasificarla llegue ahí. Esto se consigue gracias al *backpropagation* que hace que se vayan ajustando las convoluciones hasta obtener el mejor resultado.

10. **Razone y argumente a favor y en contra de usar modelos de redes CNN ya entrenados, y que se conocen han sido efectivos en otras tareas distintas de la que tiene que resolver, como modelos para aplicar directamente o como modelos a refinar para la tarea que tiene entre manos. Dar argumentos que no sean genéricos o triviales y que fundamenten su postura.**

    A favor tenemos modelos que ya han sido entrenados con millones de imágenes, se sabe que son robustos y que tienen un tasa de acierto bastante aceptable. Si el problema nuestro es distinto que el que la red soluciona, lo único que tenemos que hacer es cambiar las últimas capas de la red ya que sabemos que las primeras capas de una red son tan básicas que para este tipos de redes tan bien entrenadas podemos seguir utilizandolas. Con esto estamos ganando una red de un potencial enorme con un entrenamiento de mucho menos coste en proporción a la red que tenemos. Está claro que si el problema que tenemos que resolver está ya cubierto por una red de este tipo, nos facilita enormemente el proceso poder utilizarla.

    También, si tenemos un problema complejo en el que necesitamos una cantidad de imágenes muy grande para entrenarlo, nos conviene utilizar una red ya entrenada y modificarla si es necesario ya que así nos quitamos todo este proceso tedioso y lento.

    En contra es que si tenemos un problema "simple", puede ser que no nos merezca la pena utilizar estas redes tan complejas debido a su tamaño y entrenar nosotros una red específica con las imágenes que necesitemos y así tener un modelo más simple que resuelve el problema correctamente. Ya que si necesitamos una red que sea capaz de detectar si hay un coche, podemos entrenarla nostros solos y no utilizar una red entrenada que diferencia entre 1000 clases distintas ya que no necesitamos ese nivel de complejidad (siempre que coche no sea de esas 1000 clases ya que si no ya estaría el problema resuelto).

Referencias

[1] Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(9):891–906. Accedido el 21 de diciembre de 2018.

[2] Perona, P. (1995). Deformable kernels for early vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(5):488–499. Accedido el 21 de diciembre de 2018.

[3] Richard Szeliski (2010). Computer Vision: Algorithms and Applications Accedido el 21 de diciembre de 2018.

[4] Richard Hartley and Andrew Zisserman (2004). Multiple View Geometry in Computer Vision Accedido el 21 de diciembre de 2018.

[5] Pavel Yakimov and Vladimir Fursov (2015). Traffic Signs Detection and Tracking using Modified Hough Transform Accedido el 21 de diciembre de 2018.

[6] Alexander Shustanova and Pavel Yakimov (2017). CNN Design for Real-Time Traffic Sign Recognition Accedido el 21 de diciembre de 2018.