Universidad de Granada

Grado en Ingeniería Informática. Cuarto.

Cuestionario de teoría: 3.

Nombre de la asignatura:
Visión por Computador. Viernes de 9:30 a 11:30.

Realizado por:
Néstor Rodríguez Vico. DNI: 75573052C.
email: nrv23@correo.ugr.es

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN.

Granada, 23 de diciembre de 2017.

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.

Como bien hemos visto en clase y en la última práctica, los descriptores que usamos están basados en los gradientes de las imágenes. Por lo tanto, necesitamos que dichos gradientes estén lo suficientemente marcados como para poder realizar una correcta recuperación. También necesitamos que sean más o menos similares en las imágenes, ya que los gradientes son muy sensibles a las transformaciones proyectivas y, por lo tanto, se complica bastante el trabajo si tenemos una misma región en dos imágenes pero en una de ellas ha sufrido muchas variaciones con respecto a la otra.

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.

Hasta ahora, cuando usábamos descriptores, la idea era encontrar correspondencias entre los descriptores de la instancia a reconocer y los descriptores de la imagen en la que buscar. Con el modelo de bolsa de palabras dejamos de tener un descriptor con el que comparar, sino que comparamos con una palabra, la cual es una representación de varios descriptores. De esta forma, ganamos cierta facilidad a la hora de realizar los emparejamientos, ya que una palabra es la representación de varios descriptores similares. De esta forma, obtenemos un modelo más genérico, el cual es menos sensible a pequeños cambios, como son los cambios de perspectiva. El problema que tine es la construcción de dicha bolsa de palabras. Tal y cómo hemos visto en prácticas y en teoría, este proceso no es un proceso sencillo, en el que tenemos inconvenientes como el número de palabras que queremos obtener, sabiendo que podemos llegar a quedarnos cortos o incluso pasarnos. También es complejo decidir como extraer las características, cuantas extraer... Por lo tanto, tenemos que el modelo de bolsa de palabras depende bastante de cómo se construye esa bolsa de palabras y que parámetros se han usado para ello.

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 principal diferencia la encontramos en la propia definición de instancia y categoría. Cuando hablamos de instancia hablamos de un patrón concreto que queremos reconocer, por lo cual tenemos que saber que estamos buscando y tener un modelo que lo represente. Por ejemplo, lo que vemos en las transparencias de teoría, si tenemos un modelo que representa una lata de X marca, podemos ser capaces de comparar con ello y encontrar lo que buscamos. Sin embargo, cuando hablamos de categorías, nos referimos a una clase de objetos que, aunque representen lo mismo, no tienen porque ser parecidos, como sucede con las sillas nombradas también en los ejemplos de teoría. El problema es que definir un modelo genérico que cubra todas las posibles instancias dentro de una

categoría, no es algo fácil.

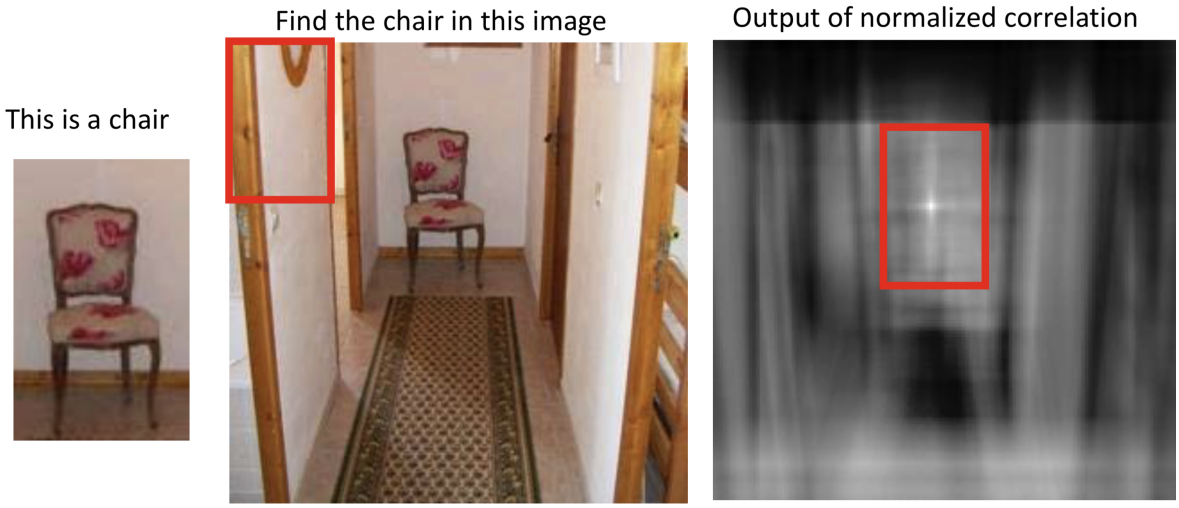
Las deformaciones que aparecen en el reconocimiento de instancias son las deformaciones de perspectiva. Por ejemplo, si estamos buscando un cubo y nos topamos con una imagen que tiene un cubo pero que ha recibido una deformación y tiene una hendidura, lo cual deforma bastante su figura, no seremos capaces de decir que ese cubo es del que buscamos. En el caso de las categorías sucede lo mismo, pero también aparece una deformación más, y es la del cambio de modelo. En el ejemplo de la silla, aparte de tener sillas deformadas, podríamos tener sillas con distinto número de patas, sillas sin respaldo, sillas de una única pieza, etcétera; tal y como podemos ver en la transparencia 11 del primer capítulo del tema 5.

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

No. Como hemos comentado antes, el reconocimiento de categoría hace referencia a reconocer multiples instancias distintas dado un modelo que representa a dicho conjunto de instancias. El problema es que el modelo no es capaz de discriminar que una palabra pertenece a una categoría u otra. Por ejemplo, supongamos que queremos diferenciar entre la categoría *coche* y la categoría *camión*. Si nos encontramos con un coche y un camión en concreto, el modelo de bolsa de palabras nos diría que hay muchas palabras en común, como pueden ser los faros o las curvas de la carrocería, y que por lo tanto en ambas imágenes aparece lo mismo. Pero, para nuestro objetivo de reconocer coches y camiones, dichas imágenes son distintas.

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.

La técnica más correcta a emplear sería la ventana deslizante. Un ejemplo de esta técnica es el que podemos ver en las transparencias de teoría:



La idea es tener una ventana que vamos paseando por la imagen en la que queremos reconocer la instancia. La idea es ir comparando la relación entre los gradientes de la instancia objetivo y los gradientes de la región cubierta por la ventana en cada momento. Podemos identificar nuestra instancia en la zona donde dicha comparación de los gradientes sea máxima. De hecho, como la ventana deslizante estará en una posición determinada en ese momento, sabremos que nuestra instancia se encuentra en esa zona de la imagen. Debemos tener cuidado con el tamaño de la ventana deslizante a usar, ya que no sabemos que tamaño tendrá la instancia en la imagen de consulta, por lo tanto lo mejor sería pasear ventanas de distinto tamaño para poder mediar con la escala.

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 seguimos enfrentándonos a un problema de reconocimiento de instancias, así que podemos seguir usando los gradientes para este trabajo. El problema que tenemos ahora es que los gradientes, y por lo tanto los descriptores, no se van a parecer en nada. Por lo tanto, en este caso usaría la técnica de bolsa de palabras, ya que esta técnica introduce más generalidad que usar los descriptores como tal. De esta forma, es más fácil encontrar correspondencias entre la instancia que estamos buscando y las imágenes donde la buscamos.

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.

El enfoque que le daría a este problema es el reconocimiento de instancias, ya que la idea es, dado una imagen, encontrar si aparece una instancia (un edificio histórico) de una conjuntos de instancias (la base de datos de edificios históricos disponibles). Para resolver este problema, usaría un modelo de bolsa de palabras junto con un fichero invertido que permita realizar la recuperación de información. Para garantizar que la recuperación es eficiente, dicho fichero invertido debe estar calculado previamente, y almacenado en el dispositivo móvil o en un servidor sobre el que realizar las consultas. De esta forma, sólo necesitaríamos realizar el proceso de consulta para la imagen proporcionada por el usuario, el cual se puede hacer en tiempo real.

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 haría lo mismo que voy a hacer para el proyecto final (algoritmo HOG para detección de peatones), que es lo mismo a lo comentado en las clases de teoría. Tendría un algoritmo que nos permita extraer una representación exhaustiva de cada imagen y luego pasar dicha representación a un clasificador que nos indique si hay o no hay señales de tráfico en dicha imagen. Por lo tanto, lo que haría sería usar, por ejemplo, el algoritmo HOG para extraer la representación de la imagen y un SVM como clasificador. Debemos tener en cuenta que, por la aplicación que tendría, necesitaríamos que la aplicación fuese en tiempo real y con unos resultados bastante fiables. Por lo tanto, el clasificador SVM debe estar entrenado previamente y almacenado para sólo tener que realizar el proceso de consulta. Otra opción sería usar una red neuronal con convolución, la cual cumple los requisitos que hemos mencionado anteriormente, nos permite hacer el proceso en tiempo real y obteniendo unos resultados bastante fiables.

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.

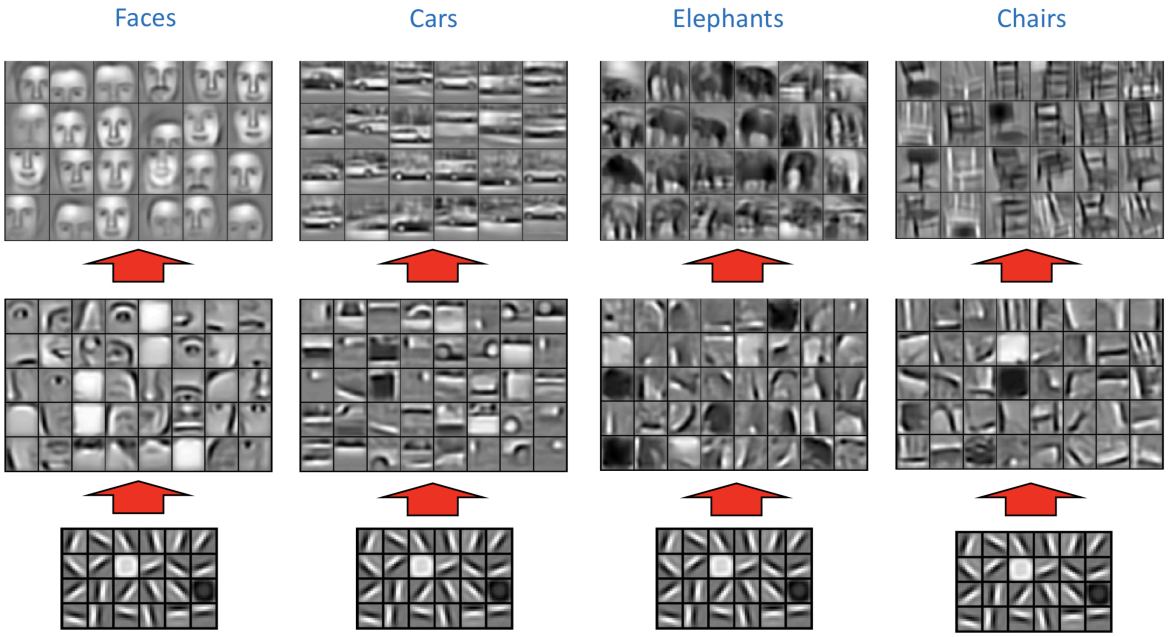
Lo más relevante que aporta los modelos CNN es el poder de generalización que tienen. La principal diferencia que tiene en cuanto a lo que se ha hecho hasta el año 2012 es el proceso de extracción de las características de una imagen. Antes debíamos definir un proceso de obtención de descriptores (nosotros nos hemos centrado en SIFT), el cuál da buenos resultados pero no siempre tiene porque darlos, ya que es un proceso fijado y que siempre es el mismo. Esto con los modelos CNN no pasa, ya que la propia red se va adaptando y corrigiendo a ella misma para ir obteniendo mejores resultados. Sin embargo, hay una idea que existía antes y sigue existiendo ahora con las redes neuronales

con convolución. Y es la idea de usar los gradientes para extraer información relevante y luego usar dicha información para procesarla, por ejemplo, en un clasificador que nos permita decidir si dicha imagen contiene un modelo que estamos buscando o no.

Mi opinión acerca del porqué han supuesto realmente una mejora es que antes no éramos capaces de extraer correctamente las características más representativas de las imágenes y ahora tenemos un modelo que lo hace el de forma dinámica, adaptando el proceso conforme va analizando los datos.

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.

En clase hemos visto que es una buena opción. Sabemos que es una buena opción porque cuando entrenamos una red neuronal con convolución, lo primero que hacemos es particionar las imágenes en pequeñas regiones para ir entrenando la red y, conforme avanzamos en la estructura de la red, estar pequeñas regiones se van combinando y formando regiones más grandes. Por lo tanto, las primera regiones que introducimos en la red neuronal son regiones muy pequeñas y que, al fin y al cabo, se pueden encontrar en cualquier imagen. Veamos esta idea:



En ambos casos, las regiones más pequeñas son similares en todas las imágenes. Por lo tanto, podemos usar una red ya entrenada (con cualquier conjunto de imágenes) ya que, en su regiones más pequeñas, serán similares a las imágenes que nosotros vayamos

a usar. Lo único que deberíamos hacer es fijar los pesos en las primeras capas de la red neuronal y reentrenarla con nuestras imágenes en las últimas capas, para adaptar y especializar la red a nuestro problema.

Pero esto es un arma de doble filo. Supongamos que la red que queremos usar no se ha entrenado con un conjunto variado de imágenes, sino que se ha entrenado con unas imágenes que cumples una determinada característica. Esto nos llevaría a tener una red demasiado especializada como para que de buenos resultados en otro campo que no sea ese. Por lo tanto, debemos usar redes que hayan sido entrenada con imágenes variadas para no tener este problema.

Bibliografía:

- Transparencias de teoría