Cómputo de Características Invariantes a la Rotación para el Reconocimiento de Distintas Clases de Objetos

Michael Villamizar, Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial, UPC-CSIC
Llorens Artigas 4-6, 08028 Barcelona, España
{mvillami, sanfeliu}@iri.upc.edu

Juan Andrade-Cetto
Centre de Visió per Computador, Universitat Autònoma de Barcelona
Edifici O, Campus UAB, 08193 Bellaterra, España
cetto@cvc.uab.es

Resumen

Este artículo presenta un sistema para el reconocimiento de objetos basado en características locales simples invariantes a escala y orientación, que al ser entrenado con un mecanismo de clasificación supervisada produce clasificadores robustos para un número limitado de clases de objetos. El sistema extrae las características más relevantes de un conjunto de muestras de entrenamiento y construye una estructura jerárquica de ellas, concentrándose primero en características comunes entre las clases y posteriormente, en aquellas características específicas de cada clase. Para lograr invariancia a rotación de forma eficiente se propone el uso de filtros orientados no Gaussianos, junto con una Imagen Integral de Orientación para un cálculo rápido de la orientación local.

1. Introducción

La detección de objetos es crucial para la mayoría de tareas de visión por ordenador; particularmente, en aplicaciones que requieren el posterior reconocimiento de estos. Los primeros enfoques para la solución del reconocimiento de objetos por ordenador se basan en la búsqueda de correspondencia entre modelos geométricos del objeto y características en la imagen. Para evitar la necesidad de poseer tales modelos a priori, surgió en las últimas dos décadas el paradigma del reconocimiento de objetos basado en apariencia, usando técnicas de reducción de dimensionalidad tales como el análisis de componentes principales. Desafortunadamente, la correspondencia basada en apariencia es propensa a fallar en situaciones con pequeñas oclusiones o bajo distintas condiciones de iluminación o del fondo. Recientemente, el reconocimiento de objetos se ha beneficiado de estrategias que combinan la combinación de características geométricas locales, así como de apariencia. El más popular de estos métodos, es quizá, el uso de descriptores SIFT [4].

A diferencia de usar reglas de saliencia para la selección de características como en el caso del descriptor SIFT, el uso de técnicas de *boosting* ha demostrado beneficios al seleccionar las características geométricas y de apariencia más relevantes sobre conjuntos de muestras de entrenamiento. A pesar de su poder para alcanzar un reconocimiento correcto con los datos del entrenamiento, las primeras técnicas de *boosting* como [9], fueron confeccionadas para el reconocimiento de una sola clase de objetos, y por consiguiente no son adecuadas para el reconocimiento multiclase, dada la gran cantidad de características que se necesitan al entrenar independientemente cada clase de objetos. Últimamente han aparecido algunas extensiones a la idea general de clasificación con *booting* que permiten el entrenamiento combinado de múltiples clases [3, 1]. En visión por ordenador, Torralba *et al.* [7] propuso una extensión al algoritmo de *boosting* (gentleboost), con el propósito de compartir características a través de múltiples clases de objetos y reducir el número total de clasificadores. Al método se le llamó *JointBoost*, y según este enfoque, todas las clases objetos son entrenadas conjuntamente, donde para cada posible subconjunto de clases ($2^n - 1$ excluyendo el conjunto fondo), la característica más útil es seleccionada para distinguir aquel subconjunto de la clase fondo. El proceso es repetido hasta que el error global de clasificación alcanza un mínimo, o hasta alcanzar un número limite de clasificadores.

El tipo de clasificadores usados en [7] están basados en la correspondencia de simples patrones, que presumi-

Este trabajo es posible gracias al *EURON Network Robot Systems Research Atelier* NoE-507728 y al proyecto NAVROB DPI 2004-05414 del Ministerio Español de Educación y Ciencia, para M. Villamizar y A. Sanfeliu. J. Andrade-Cetto es investigador postdoctoral Juan de la Cierva del Ministerio Español de Educación y Ciencia bajo el proyecto TIC2003-09291, y es también patrocinado en parte por el proyecto europeo PACO-PLUS FP6-2004-IST-4-27657. Los autores pertenecen al Grupo de Investigación Consolidado Visión Artificial y Sistemas Inteligentes de la Dirección General de Investigación de Cataluña.

blemente fallaría si el objeto se encuentra a diferentes orientaciones de las cuales fue entrenado. En este trabajo se investiga sobre la selección de características multiclase de forma similar, pero con un fuerte interés en el cálculo rápido de clasificadores débiles invariantes a orientación [8], con el fin de obtener un sistema de reconocimiento de múltiples objetos que sea invariante a rotación.

En [9], Viola introdujo la Imagen Integral para la evaluación rápida de características. Una vez calculada, la Imagen Integral permite la respuesta de una imagen a la convolución con una base de Haar [5] a cualquier posición y escala en tiempo real. Desafortunadamente, tal sistema no es invariante a cambios de orientación del objeto ni a oclusiones. Otros sistemas de reconocimiento que funcionan bien en escenas complejas son los basados en el cómputo de características locales multi-escala tales como el descriptor SIFT [4] mencionado anteriormente. Una idea importante en el descriptor SIFT es que incorpora orientación local para cada punto de interés, permitiendo invariancia a escala y orientación durante el reconocimiento. Incluso, cuando un gran número de características SIFT pueden ser calculadas en tiempo real para una única imagen, la correspondencia entre la muestra y las imágenes de prueba es realizada mediante la búsqueda del vecino más cercano y por votación con la transformada de Hough generalizada, seguido por la solución de la relación de afinidad entre vistas, lo cual podría terminar en un proceso computacionalmente costoso.

Yokono and Poggio [10, 11] prefieren el uso de esquinas *Harris* a varios niveles de resolución como puntos de interés, y de estas, seleccionan como características del objeto aquellas que son más robustas a filtros de derivadas de la Gausiana bajo rotación y escala. Como las derivadas de la Gausiana no son invariantes a rotación, ellos utilizan filtros orientados [2] para orientar todas las respuestas de las características según la orientación local de gradientes alrededor del punto de interés. En la fase de reconocimiento, el sistema aún requiere correspondencia de características locales, e iterar sobre todos las correspondencias en grupos de 6, buscando la mejor homografía, usando RANSAC para eliminar valores atípicos. Desafortunadamente, el coste computacional de este enfoque no se reporta.

En [8] hemos reportado que la respuesta del filtro a convoluciones con bases de Haar no solo se puede calcular eficientemente con una Imagen Integral, sino que puede ser aproximadamente rotada con algunas simplificaciones de los filtros orientados de Gausianas. Lo cual permite el cálculo eficiente de la respuesta a filtros invariantes a rotaciones, y su uso como clasificadores débiles.

En este artículo, hemos incorporado estas dos ideas, *boosting* multiclase e invariancia a rotación, para la selección de características comunes y específicas para construir una estructura jerárquica que permita reconocer múltiples objetos independientemente de su posición, escala y orientación utilizando un reducido conjunto de características. En nuestro sistema, los puntos de interés son seleccionados como aquellas regiones en la imagen que tienen la respuesta más discriminante bajo la convolución con un conjunto de funciones base *wavelets* a varias escalas y orientaciones. En la Sección 2 se explica como las características más relevantes son seleccionadas y combinadas para clasificar múltiples objetos. La selección se basa en *JointBoost*, en el cual una estructura jerárquica es formada por conjuntos de clasificadores comunes y específicos. Una combinación lineal de estos clasificadores débiles produce un clasificador fuerte para cada clase de objetos, el cual es usado en la fase de detección. La invariancia a rotación es posible al convolucionar cada muestra con funciones base orientadas. La rotación del filtro se calcula de forma eficiente con la ayuda de los filtros orientados, que es a su vez la combinación lineal de filtros base, tal y como se indica en la Sección 3.

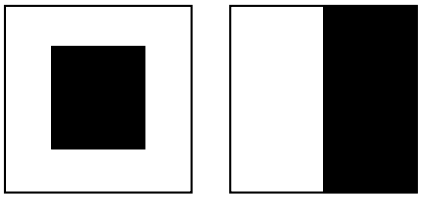
Durante la fase de reconocimiento, cada región de la imagen debe ser rotada a una orientación canónica, obtenida durante el entrenamiento, antes de hacer la correspondencia. Tal orientación es dictada por la moda del histograma de orientaciones de gradientes descrito en la Sección 4. En la Sección 5 se explica nuestra Imagen Integral de orientaciones, la cual proponemos para un cálculo rápido de orientación, y en la sección 6 se presentan algunos experimentos.

2. Selección de Características

El conjunto de características locales que mejor discriminan un objeto se obtiene gracias a la convolución de muestras de imágenes positivas con un conjunto simplificado de funciones bases *wavelets* [5] a diferentes escalas y orientaciones. Estos filtros tienen selectividad a la orientación espacial como también a frecuencia, y producen características que capturan el contraste entre regiones representando contornos, puntos y líneas. El conjunto de operadores usados se muestra en la Figura 1. La respuesta del filtro es equivalente a la diferencia de intensidad en la imagen original entre la región oscura y la clara. La figura 1 d) ilustra como un objeto puede ser representado por un pequeño conjunto de características relevantes.

La convolución de estos operadores a una orientación deseada se realiza al orientar el filtro (Sección 3), y la convolución rápida sobre cualquier región de la imagen original es eficientemente obtenida mediante la Imagen Integral (Sección 5).

La selección de características se lleva a cabo de man-



(a) f_0 (b) f_1

(c) f_2 (d)

Figura 1: Conjunto de bases *wavelet* simplificado. a) punto b) contorno, c) línea, y d) características locales del objeto

era similar a *JointBoost* [7], escogiendo el clasificador débil $h(I, s)$ que mejor discrimina cualquier subconjunto s de la clase fondo, con $2n - 1$ subconjuntos de clases $c = 1...n$ (excluyendo el mismo conjunto fondo), El clasificador débil se define por los parámetros del filtro, tamaño, posición, orientación y umbral, tomando valores de decisión binaria

$$h(I, s) = \begin{cases} 1 & : \quad I * f > t \\ 0 & : \quad \text{otros casos} \end{cases} \tag{1}$$

donde I es una imagen de entrenamiento perteneciente a la clase c en el subconjunto s, f es el filtro con todos sus parámetros, $*$ indica la operación de convolución, y t es el umbral de la respuesta del filtro.

En cada iteración durante la fase de entrenamiento, el algoritmo debe buscar para todos los $2n - 1$ subconjuntos, el clasificador débil que mejor discrimine aquel subconjunto de la clase fondo, minimizando el error cuadrático sobre las muestras ponderadas de todas las clases en aquel subconjunto.

$$J_{wse} = \sum_{c=1}^{n} \sum_{s=1}^{m} w_i^c (z_i^c - h(I, s))^2 \tag{2}$$

donde z_i^c y w_i^c son la etiqueta y el peso de la muestra i para la clase c respectivamente, y m el número total de muestras de entrenamiento. El algoritmo también actualiza los conjuntos de pesos sobre las muestras de entrenamiento. El número de conjuntos corresponde con el número de clases a aprender. Inicialmente, todos los pesos son iguales, pero en cada iteración, los pesos de las muestras mal clasificadas son incrementados para que el algoritmo sea forzado a concentrarse en las muestras difíciles del conjunto de entrenamiento que han sido mal clasificadas por los anteriores clasificadores. Finalmente, se selecciona el clasificador débil para el subconjunto que tiene el mínimo error cuadrático J, y se añade iterativamente al clasificador fuerte para cada clase c en s, $H(I, c)$.

$$H(I, c) := H(I, c) + h(I, s) \tag{3}$$

La invariancia a escala se obtiene al iterar también sobre filtros escalados. El escalado del los filtros se puede efectuar en tiempo constante con la Imagen Integral computada previamente.

3. Filtros Orientados

Para obtener invariancia a orientación, los filtros locales deben ser rotados antes de realizar la convolución. Una buena alternativa, es computar estas rotaciones con filtros orientados [2], o con su versión compleja [6]. Un filtro orientado es un filtro rotado a partir de la combinación lineal de un conjunto de filtros base a orientaciones específicas.

$$I * f(\theta) = \sum^{n} k_i(\theta) I * f(\theta_i) \ , \tag{4}$$

donde $f(\theta_i)$ son los filtros base orientados, y k_i son los coeficientes de las bases.

Considere por ejemplo, la función Gausiana $G(u, v) = e^{-(u^2+v^2)}$, y su primera y segunda derivadas $G'_u = -2ue^{-(u^2+v^2)}$ y $G''_u = (4u^2 - 2)e^{-(u^2+v^2)}$. Estos filtros pueden ser orientados como la combinación lineal de los filtros base. El número de filtros base es uno más que el orden de diferenciación.

Consecuentemente, la primera derivada de la función Gausiana a cualquier orientación θ es

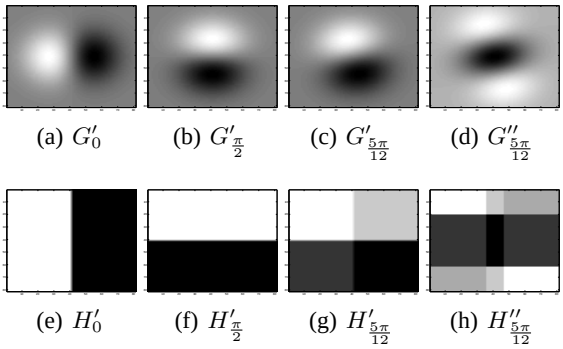
$$G'_\theta = \cos\theta G'_u + \sin\theta G'_v \ , \tag{5}$$

y, la orientación de la segunda derivada de la Guasiana puede ser obtenida con

$$G''_\theta = \sum_{i=1}^{3} k_i(\theta) G''_{\theta_i} \tag{6}$$

con $k_i(\theta) = \frac{1}{3}(1 + 2\cos(\theta - \theta_i))$; y G''_{θ_i} son filtros precalculados de segunda derivada a $\theta_1 = 0$, $\theta_2 = \frac{\pi}{3}$, y $\theta_3 = \frac{2\pi}{3}$. Ver Figura 2.

El convolucionar con filtros de Gausiana es un proceso computacionalmente costoso. En cambio, hemos propuesto en [8] que es posible aproximar la respuestas de tales filtros por la de filtros con bases Haar usando la Imagen Integral, lo cuál es computacionalmente más eficiente. Así, aproximamos la respuesta de la primera derivada orientada con



(a) G'_0 (b) $G'_{\frac{\pi}{2}}$ (c) $G'_{\frac{5\pi}{12}}$ (d) $G''_{\frac{5\pi}{12}}$

(e) H'_0 (f) $H'_{\frac{\pi}{2}}$ (g) $H'_{\frac{5\pi}{12}}$ (h) $H''_{\frac{5\pi}{12}}$

Figura 2: Filtros orientados de primer y segundo orden. (a-b) bases Gausianas, (c-d) filtros orientados de la Gausiana, (e-f) bases de Haar, (g-h) filtros Haar orientados.

$$I * f_1(\theta) = \cos\theta I * f_1(0) + \sin\theta I * f_1(\tfrac{\pi}{2}) \ . \quad (7)$$

y de la misma forma, el detector de líneas a cualquier orientación θ se obtiene con

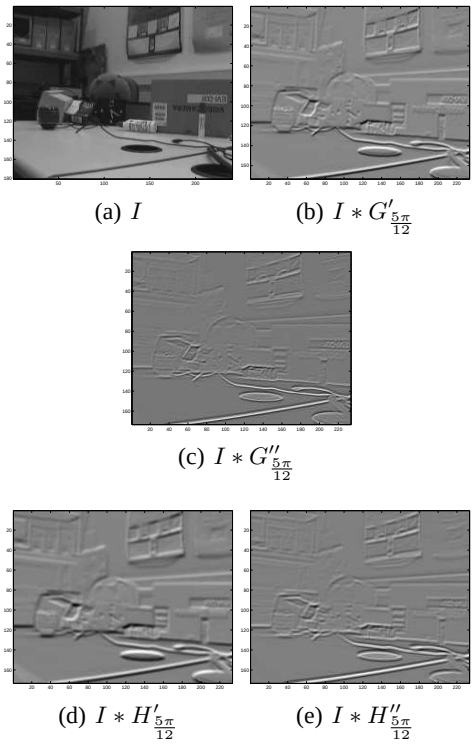
$$I * f_2(\theta) = \sum_{i=1}^{3} k_i(\theta) I * f_2(\theta_i) \ . \quad (8)$$

La similitud de la respuesta a filtros con bases Gausianas y de Haar nos permite usar esta última base como clasificadores débiles para la detección de puntos, contornos y líneas, de igual forma que se podría hacer con los filtros Gausianos. La principal ventaja de este enfoque es la velocidad de computo. Mientras que la convolución con filtros Gausianos tiene una complejidad computacional de $O(n)$, con n el tamaño del filtro, la convolución con las bases de Haar orientadas puede ser calculada en tiempo constante usando la representación de Imagen Integral. La Figura 3 muestras algunos resultados.

4. Orientación Local

Considere que una sesión de entrenamiento ha producido una constelación H de características locales h como se ilustra en la Figura 4. El objetivo es examinar para múltiples posiciones y escalas en cada nueva imagen, si tal constelación pasa la prueba H o no. Envés de intentar cada posible orientación de la constelación, escojemos almacenar la orientación canónica θ_0 de H de una imagen de referencia del conjunto de entrenamiento, y compararla con la orientación θ de cada región de la imagen que está siendo examinada. La diferencia de orientación entre las dos indica la cantidad que se debe reorientar la constelación de características antes de examinar contra el clasificador H.

Una forma para calcular la orientación de una región es con la razón de las primeras derivadas de la Gausiana G'_u y G'_v [11], $\tan\theta = \frac{I*G'_v}{I*G'_u}$. Otra técnica, más robusta a oclusiones parciales es usar la moda del histograma de orientaciones de los gradientes locales (ver Figura 4 c-d), para lo cual es necesario calcular la orientación del gradiente pixel a pixel, envés de una convolución con toda la región como en el caso previo.



(a) I (b) $I * G'_{\frac{5\pi}{12}}$

(c) $I * G''_{\frac{5\pi}{12}}$

(d) $I * H'_{\frac{5\pi}{12}}$ (e) $I * H''_{\frac{5\pi}{12}}$

Figura 3: Respuesta a filtros de primer y segundo orden. a) imagen original , b-e) respuestas de los filtros

5. Imagen Integral de Orientación

Una Imagen Integral es una representación de la imagen que permite un cálculo rápido de características, debido a que no trabaja directamente con las intensidades de la imagen original. En cambio, esta trabaja sobre una imagen incrementalmente construida que añade valores de características a lo largo de filas y columnas. Una vez computada esta representación de la imagen, cualquiera de nuestras características locales (clasificador débil) puede ser computado en cualquier localización y escala en tiempo constante

En su forma más simple, el valor de la Imagen Integral M en la coordenada u, v contiene la suma de los valores de los pixeles superior e izquierda de u, v, incluyendo estos últimos.

$$M(u,v) = \sum_{i \le u, j \le v} I(i,j) \quad (9)$$

Entonces, es posible computar por ejemplo, la suma de valores de intensidades en una región rectangular

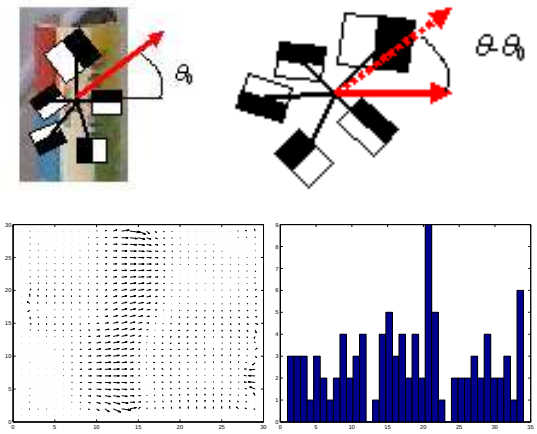


Figura 4: Orientación local. a) orientación canónica, b) constelación rotada, c) gradientes de Imagen, b) histograma de orientaciones de gradientes.

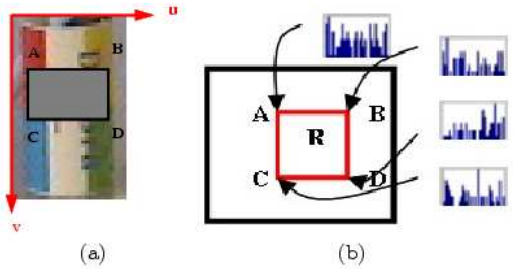


Figura 5: Imágenes integrales, a) imagen integral b) imagen integral de orientaciones.

simplemente añadiendo y sustraendo las intensidades acumuladas en sus cuatro vértices en la Imagen Integral(Figura 5a). Luego, la respuesta de filtros Haar puede ser calculada de forma rápida independientemente del tamaño y posición del filtro.

$$\text{Area} = A + D - B - C \tag{10}$$

Extendiendo la idea de tener información acumulada en cada pixel en la Imagen Integral, hemos decidido almacenar en ella información de histogramas de orientación envés de intensidades acumuladas. Una vez construida esta Imagen Integral de Orientaciones, es posible computar el histograma de orientación local para cualquier región rectangular dentro de la imagen en tiempo constante. Ver Figura 5b.

$$\begin{aligned}\text{Histograma(Area)} = \text{Histograma}(A) + \text{Histograma}(D) \\ -\text{Histograma}(B) - \text{Histograma}(C)\end{aligned} \tag{11}$$

6. Experimentos

En este artículo reportamos algunos resultados iniciales para un número limitado de objetos en imágenes

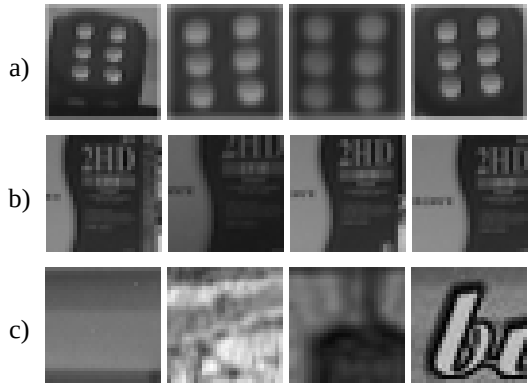


Figura 6: Clases de objetos a entrenar. a) imágenes del dado, b) imágenes de la caja de discos , y c) imágenes del fondo.

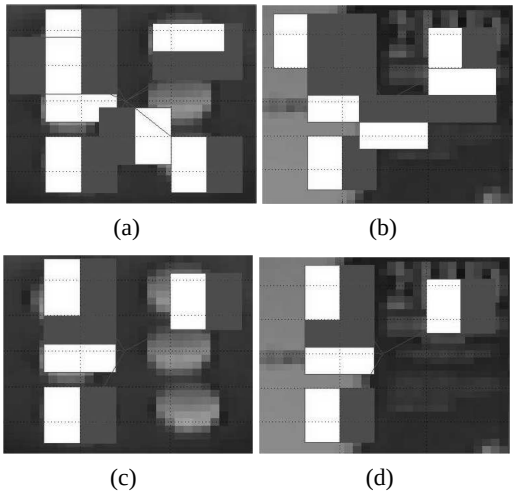
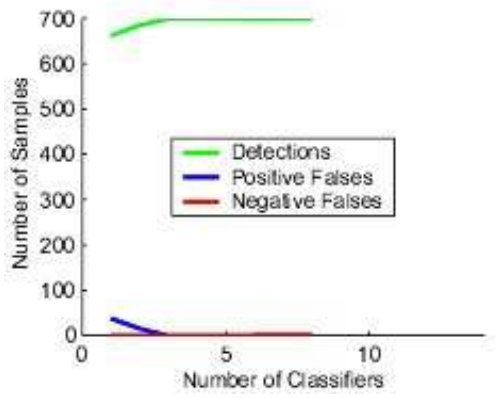


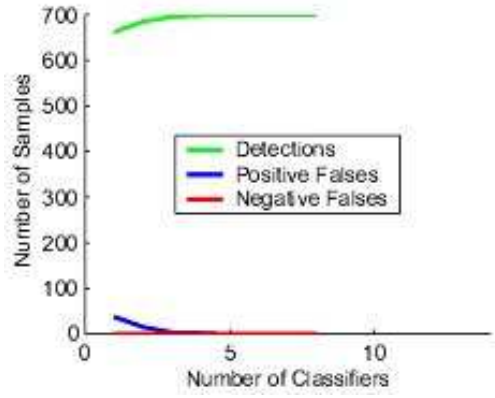
Figura 7: Constelaciones. a) constelación del dado b) constelación de la caja de discos (c-d) clasificadores comunes.

en escala de grises. El conjunto de entrenamiento tenía 100 imágenes para cada clase, y 500 imágenes negativas o de fondo. Estas imágenes negativas fueron extraídas de escenas exteriores e interiores. Las imágenes de los objetos de cada clase usadas para el entrenamiento presentaban algunas pequeñas trasformaciones como traslación, orientación y escala, como se observa en la Figura 6.

La Figura 7 a) y b) muestra ejemplos de constelaciones de características extraídas para cada objeto de las distintas clases. Cada una está formada por 8 clasificadores débiles (características Haar), con 4 de ellas comunes para ambas clases, y las 4 restantes específicas para cada clase. Así, produciendo una estructura jerárquica de clasificadores débiles. Imágenes c) y d) muestran solo estos 4 clasificadores comunes. Ellos capturan información local similar en ambas clases, separando estos de la clase fondo, sin la necesidad de ser específica.



(a)



(b)

Figura 8: Desempeño en el entrenamiento. a) dado b) caja de discos.

El clasificador fuerte puede ser expresado como la combinación de clasificadores débiles comunes y específicos. Considere el dado como la clase 1, la caja de discos la clase 2, y c_{12} el conjunto de muestras de entrenamiento conteniendo uno o ambos objetos. Entonces,

$$H(I, c_1) = \sum h(I, c_{12}) + \sum h(I, c_1) \qquad (12)$$

$$H(I, c_2) = \sum h(I, c_{12}) + \sum h(I, c_2) \qquad (13)$$

Las curvas de entrenamiento se muestran en la Figura 8. En ellas se muestra como la correcta clasificación durante el entrenamiento es alcanzada. Algunos resultados en el proceso de detección sobre una secuencia de imágenes se muestran en la Figura 9.

7. Conclusiones

En este artículo se ha introducido una selección de estructura jerárquica de características que reduce el número total de clasificadores débiles necesarios para detectar múltiples clases de objetos. Con este método

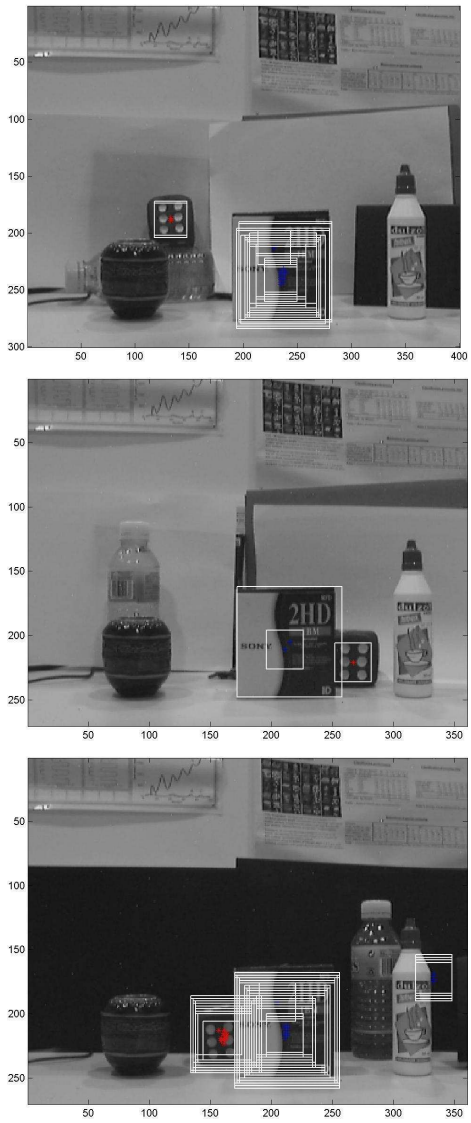


Figura 9: Ejemplos de correcta detección de clasificadores entrenados conjuntamente (dado y caja de discos). La última imagen muestra también bajo que circunstancias puede ocurrir una falsa detección.

el sistema encuentra características comunes entre objetos y generaliza el problema de reconocimiento.

Nuestro enfoque se basa en seleccionar de forma automática un conjunto de características locales. En contraste con enfoques previos, hemos propuesto el uso de filtros con bases de Haar y una nueva imagen integral de orientación para la evaluación rápida de la orientación local.

Referencias

[1] G. Eibl and K-P. Pfeiffer. Multiclass boosting for weak classifiers. *J. Mach. Learn. Res.*, 6:189–210, 2005.

[2] W. T. Freeman and E. H. Adelson. The design

and use of steerable filters. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(9):891–906, 1991.

[3] L. Li. Multiclass boosting with repartitioning. In *Proc. 23rd Int. Conf. Machine Learning*, Pittsburgh, 2006. To appear.

[4] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[5] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. IEEE Int. Conf. Comput. Vision*, page 555, Bombay, Jan. 1998.

[6] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. 7th European Conf. Comput. Vision*, pages 414–431, Copenhagen, 2002. Springer-Verlag.

[7] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. 18th IEEE Conf. Comput. Vision Pattern Recog.*, pages 762–769, Washington, Jul. 2004.

[8] M. Villamizar, A. Sanfeliu, and J. Andrade-Cetto. Computation of rotation local invariant features using the integral image for real time object detection. In *Proc. 18th IAPR Int. Conf. Pattern Recog.*, Hong Kong, Aug. 2006. IEEE Comp. Soc. To appear.

[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 15th IEEE Conf. Comput. Vision Pattern Recog.*, pages 511–518, Kauai, Dec. 2001.

[10] J.J. Yokono and T. Poggio. Oriented filters for object recognition: An empirical study. In *Proc. 6th IEEE Int. Conf. Automatic Face Gesture Recog.*, pages 755–760, Seoul, 2004.

[11] J.J. Yokono and T. Poggio. Rotation invariant object recognition from one training example. Technical Report 2004-010, MIT AI Lab., Apr. 2004.