Estudio Experimental sobre la Combinación Temporal de Resultados en el Reconocimiento de Caras con Secuencias de Video

O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández
Universidad de Las Palmas de Gran Canaria
Departamento de Informática y Sistemas
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas - España
{odeniz,mcastrillon,jlorenzo,mhernandez}@dis.ulpgc.es

Resumen: *Se ha comprobado que la combinación de los resultados de clasificación de varias imágenes de una secuencia mejora la probabilidad de acierto en el problema del reconocimiento de caras. No obstante, queda por dilucidar qué método de agregación temporal de los resultados es el más apropiado para cada caso concreto. En sistemas prácticos el método de combinación debe además ser simple para no consumir mucho tiempo de cómputo, teniendo en cuenta que el sistema tendrá otras etapas de proceso con una latencia relativamente alta. En este trabajo se describe un estudio experimental de varios métodos de combinación. Si bien en condiciones normales se observa que los métodos de combinación con el máximo y con la media producen los mejores resultados, la introducción de errores simulados de clasificación de las imágenes de la secuencia permite concluir que el método más robusto es el de la mayoría de votos. El costo computacional de los métodos analizados es prácticamente equivalente, con lo que se propone la regla de la mayoría de votos como el método más adecuado de entre los estudiados para su utilización en sistemas prácticos.*

Palabras clave: visión por computador, reconocimiento de caras, fusión temporal, combinación temporal

1 Introducción

A pesar de su complejidad, el problema del reconocimiento de caras ha despertado un enorme interés investigador en los últimos años. Tanto es así que algunos sistemas permiten lograr ya un rendimiento aceptable en condiciones restringidas. No obstante,

en muchos de los trabajos aparecidos se realizan experimentos en condiciones de funcionamiento en cierto sentido alejadas de la realidad: se usan bases de datos con pocas imágenes y la comparación se hace generalmente con una sola imagen. En un sistema real dispondríamos de una o más cámaras de video que proporcionarían secuencias de imágenes. El sistema de reconocimiento trataría de proporcionar en todo momento la identidad más probable de la persona que está frente a la cámara. Está demostrado que el uso de la información que proporciona la secuencia de imágenes mejora notablemente la respuesta del sistema, si se compara con el uso de una sola imagen. En este sentido, algunos de los sistemas descritos se basan en modificar la representación o los clasificadores utilizados para tener en cuenta la información que proporcionan las secuencias. Otra opción empleada es realizar simplemente una fusión de los resultados de clasificación. Con respecto a los primeros, en [1] se caracterizan las caras mediante trayectorias o caminos en el espacio de representación, obtenidas por ejemplo a partir de secuencias en las que el sujeto rota la cabeza a intervalos regulares. El reconocimiento se hace comparando la trayectoria correspondiente a la secuencia de test con las de secuencias prototipo. En [2] se forma un subespacio de representación con la secuencia de imágenes de entrada, que se compara con subespacios obtenidos en la fase de entrenamiento, apreciándose un aumento de la robustez del sistema ante cambios de expresión y de pose. En el problema de la verificación, en [3] se muestra como el uso de múltiples imágenes permite reducir el error hasta en un 40%. Asimismo, se observa que la reducción en el error es mayor al principio, pero que tiende a saturarse con el número de imágenes acumuladas. En [4] se enfatiza que el orden relativo de las imágenes de la secuencia constituye una valiosa información para el reconocimiento. Se utiliza una representación que además de las imágenes incluye información sobre el orden en que aparecen, representación que se obtiene con redes neuronales *recurrentes*. Con respecto a los trabajos en los que se realiza una fusión de los resultados, no impera un método común de fusión o combinación, siendo los más utilizados la regla del máximo [5, 6], la media [7] o la suma [8].

Si bien son posibles muchos otros esquemas de fusión, en este trabajo se realiza un estudio experimental centrado en los métodos más empleados en sistemas prácticos. Los métodos estudiados tienen todos un bajo costo computacional, restricción necesaria para su utilización en sistemas con tiempo de respuesta razonable. En la Seccion 2 se describe el sistema utilizado para obtener las secuencias empleadas en los experimentos. En la Sección 3 se describen brevemente los métodos a estudiar mientras que en la Sección 4 se comentan los resultados experimentales obtenidos. Por último, en la Sección 5 se detallan las conclusiones más importantes.

2 DESEO

Las secuencias utilizadas en los experimentos se obtuvieron con el sistema DESEO (Detección y Seguimiento de Objetos) [9]. DESEO es un sistema hardware-software capaz de realizar detección y seguimiento de personas en tiempo real, empleando información de movimiento y/o color de la piel. Las imágenes que proporciona DESEO son procesadas

para confirmar que realmente estamos ante una cara y, si es así, normalizarla para su posterior reconocimiento. En esta etapa se ajusta una elipse a la mayor región detectada en la imagen con color de la piel. Con los parámetros de esta elipse ya podemos descartar heurísticamente valores que no corresponden a caras frontales, que son las que más interesan para el reconocimiento. También se utilizan los parámetros para hacer una rotación con el fin de dejar la cara vertical. A continuación, utilizando una transformada de simetría se realiza una búsqueda de la posición de los ojos (el operador de simetría tiene un alto costo computacional, con lo que la búsqueda se realiza en una determinada zona con respecto a la elipse). La probable posición de los ojos permite descartar imágenes que no presenten valores coherentes, como por ejemplo una distancia entre ojos muy pequeña con respecto a la elipse. La distancia entre ojos se utiliza para escalar la imagen a un tamaño fijo. El proceso completo se describe con más detalle en [10]. La mayoría de los sistemas de detección y seguimiento de personas para el posterior análisis de las caras se basan en técnicas parecidas. El resultado neto es un conjunto de imágenes de la cara, normalizadas y listas para su reconocimiento, ver Figura 1.

Figura 1. Dos secuencias de imágenes faciales normalizadas, obtenidas con el sistema descrito en la Sección 2, y utilizadas en los experimentos.

El sistema tarda generalmente un tiempo entre 100-200 ms en procesar cada frame, para un tamaño de región de piel de 60x85, usando un Celeron a 433 Mhz. De ahí la importancia de que el tiempo de combinación sea mínimo. Es necesario señalar que un sistema como el descrito no logrará en todos los casos detectar las imágenes en las que aparece la cara frontal. Del mismo modo, siempre se producirá una cierta tasa de falsos positivos, como los que aparecen en la Figura 2.

Figura 2. Ejemplos de fallos de los procesos de detección y normalización de las caras con el sistema descrito en la Sección 2.

3 Métodos de combinación

Los métodos de combinación empleados en los experimentos se dividen en dos tipos. Por un lado, métodos que hacen uso de estimaciones de la probabilidad de clase a posteriori: Media, Mayoría de votos, Máximo y Jurado. La regla de la media consiste en hallar la media de las probabilidades a posteriori de cada clase, y asignar a la secuencia la identidad de la clase con mayor valor de esta probabilidad media. Para el cálculo de la media, se utilizó la relación $(n-1)/n*Media_{n-1}+p_c^n/n$, donde n es el número de imágenes consideradas. La regla del máximo utiliza el máximo de las probabilidades a posteriori. El método del jurado [11] consiste en descartar las salidas del clasificador más alta y más baja, y mediar el resto. Por otro lado, los métodos basados en el orden de clasificación mantienen una lista con el orden de pertenencias a clase que da el clasificador. Esto es, para una imagen dada, la primera clase será la que recibe mayor probabilidad a posteriori del clasificador, la segunda la siguiente en mayor probabilidad y así sucesivamente. Los métodos basados en el orden de clasificación que se emplearon son Posición más alta y Cuenta Borda [12]. El primero adjudica la decisión a la clase que aparece en una posición más alta en la lista. Con cada nueva clasificación, el método de la cuenta Borda suma las nuevas posiciones a las anteriormente obtenidas.

4 Resultados de los experimentos

Los experimentos se realizaron con secuencias de imágenes obtenidas con el sistema descrito en el apartado 2. Se utilizaron 13 secuencias, una por individuo, cada una de 167 imágenes, 2171 imágenes distintas en total. Las imágenes tienen un tamaño de 39x43 pixels. De cada secuencia, se tomaron 3 imágenes de entrenamiento, 39 en total, y el resto para test. Sobre el conjunto de las imágenes de entrenamiento se aplicó la técnica PCA *(Principal Component Analysis)*, donde el número K de coeficientes retenidos se escogió como aquel que daba un menor error en el conjunto de test. Cada experimento se realizó diez veces, cambiando cada vez el orden de las imágenes de cada secuencia. Los resultados finales son la media de los diez resultados parciales. En la Figura 4 aparece la tasa de acierto de los distintos métodos. Las imágenes de test de cada secuencia se tomaron de n en n, con solapamiento (ver Figura 3), donde n es el número de imágenes acumuladas que va de 1 a 30 [1]. Como clasificador se utilizó en un caso el vecino más cercano (utilizando la media como prototipo y distancia euclídea) y en otro un clasificador basado en SVM (*Support Vector Machines*), con kernel de función de base radial.

En la Tabla I se muestra como medida numérica de comparación entre los métodos el área bajo la curva en la Figura 4, así como el máximo alcanzado habiendo considerado el total de 30 imágenes. El método del oráculo se incluye como medida comparativa. Con el método del oráculo, una vez que con una de las imágenes de la secuencia el clasificador da la identidad correcta, se mantiene esa identidad independientemente de resultados

[1] En el sistema descrito en la Sección 2 los 30 frames se clasificarían en un tiempo entre 3 y 6 segundos, lo que constituye una cota superior de tiempo de respuesta para un sistema práctico.

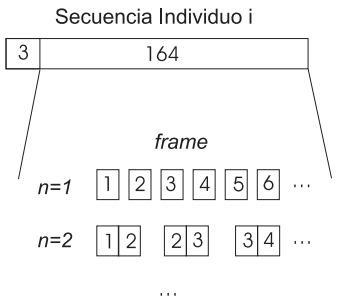

Figura 3. Formación de las secuencias utilizadas en los experimentos.

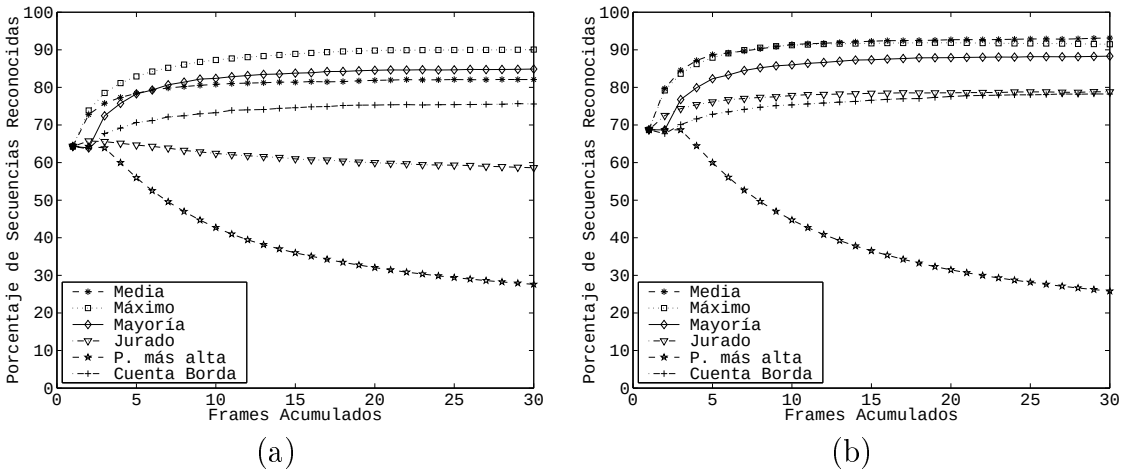

Figura 4. Porcentaje de secuencias de test acertadas en función del número de frames acumulados para los distintos métodos de combinación estudiados, (a) usando el clasificador de media más cercana con distancia euclídea, (b) utilizando un clasificador SVM con kernel de función de base radial y $\sigma = 1400$.

posteriores.

De los resultados expuestos se aprecia que el máximo y la media son los métodos de combinación más efectivos. La superioridad de la combinación con el máximo resultó estadísticamente significativa para el caso del vecino más cercano. La principal medida a tener en cuenta es en este caso el área bajo la curva, aunque el máximo alcanzado también es importante. En contraste con los resultados obtenidos por otros autores para el problema de combinación de clasificadores, los métodos basados en el orden de clasificación no dan buenos resultados para la aplicación de estudio, aunque como se verá más adelante, hay casos en los que resultan mejores que otros métodos. Independientemente de esto, puede verse un efecto de saturación en la combinación de los resultados de clasificación, descrito ya en [3].

Es necesario reseñar que el sistema descrito en la Sección 2 funciona relativamente bien para el conjunto de imágenes con las que se probó, pero que aún no ha sido probado

Tabla I. Valores de área bajo la curva (AUC) y máxima tasa de aciertos alcanzado (MAX) para los distintos métodos de combinación, para secuencias de hasta 30 frames, utilizando el clasificador de vecino más cercano (VMC) y el clasificador basado en SVM.

	VMC		SVM	
	AUC	MAX	AUC	MAX
Media	2331.06	82.81 %	2630.28	93.33 %
Mayoría de votos	2371.49	85.16 %	2477.48	88.52 %
Máximo	2521.97	90.28 %	2611.18	92.96 %
Jurado	1782.24	66.36 %	2250.92	79.45 %
Posición más alta	1151.35	64.96 %	1177.39	69.61 %
Cuenta Borda	2128.83	76.06 %	2197.09	78.84 %
Oráculo	2780.40	98.68 %	2820.69	99.66 %

en condiciones de funcionamiento real, donde los errores son numerosos. El sistema utiliza numerosos criterios heurísticos determinados empíricamente, y además la detección por color no es muy robusta ante cambios de iluminación. Por ello, en funcionamiento real es de esperar un aumento de los falsos positivos en la detección de caras frontales, como los ilustrados en la Figura 2. Estos falsos positivos se traducirán generalmente en resultados de clasificación erróneos. Con el fin de comprobar el efecto de este aspecto en los distintos métodos de combinación, se introdujeron errores en la salida proporcionada por el clasificador. El resultado de clasificación de un cierto porcentaje de las imágenes de test, seleccionadas al azar, se hizo aleatorio. Por otro lado, teniendo en cuenta que los errores de este tipo aparecerán generalmente en frames consecutivos de la secuencia, se introdujo un parámetro adicional. El parámetro mide el grado de cercanía (GC) en los fallos introducidos. La distribución se realiza en base a una normal cuya varianza depende del parámetro. Cuando GC tiende al 100 %, la varianza tiende a 0 y cuando GC tiende a 0 % la varianza tiende a infinito. En un extremo los fallos se distribuyen al azar, mientras que en el otro se distribuyen consecutivos. La distribución de los fallos al azar correspondería a fallos producidos en los niveles más bajos del procesamiento, como la etapa de adquisición de imágenes, mientras que los fallos consecutivos corresponden generalmente a situaciones en las que fallan los niveles superiores. La respuesta de los distintos métodos ante el error introducido se muestra en las Figuras 5 y 6 como la variación de área bajo la curva de cada método, para cierta cantidad de error. Puede observarse que el método más robusto es el de la mayoría de votos. La combinación con el máximo pasa a ser uno de los peores métodos. Además, puede observarse que la mayoría de votos pasa a ser el mejor método incluso cuando el error introducido es pequeño. Con respecto al grado de cercanía de los errores introducidos, se observa que el método de la mayoría es cada vez mejor a medida que los errores son más dispersos.

La importancia del error se manifiesta claramente cuando se tiene una secuencia en la que aparecen dos personas. Si los niveles inferiores no se diseñan para detectar los cambios

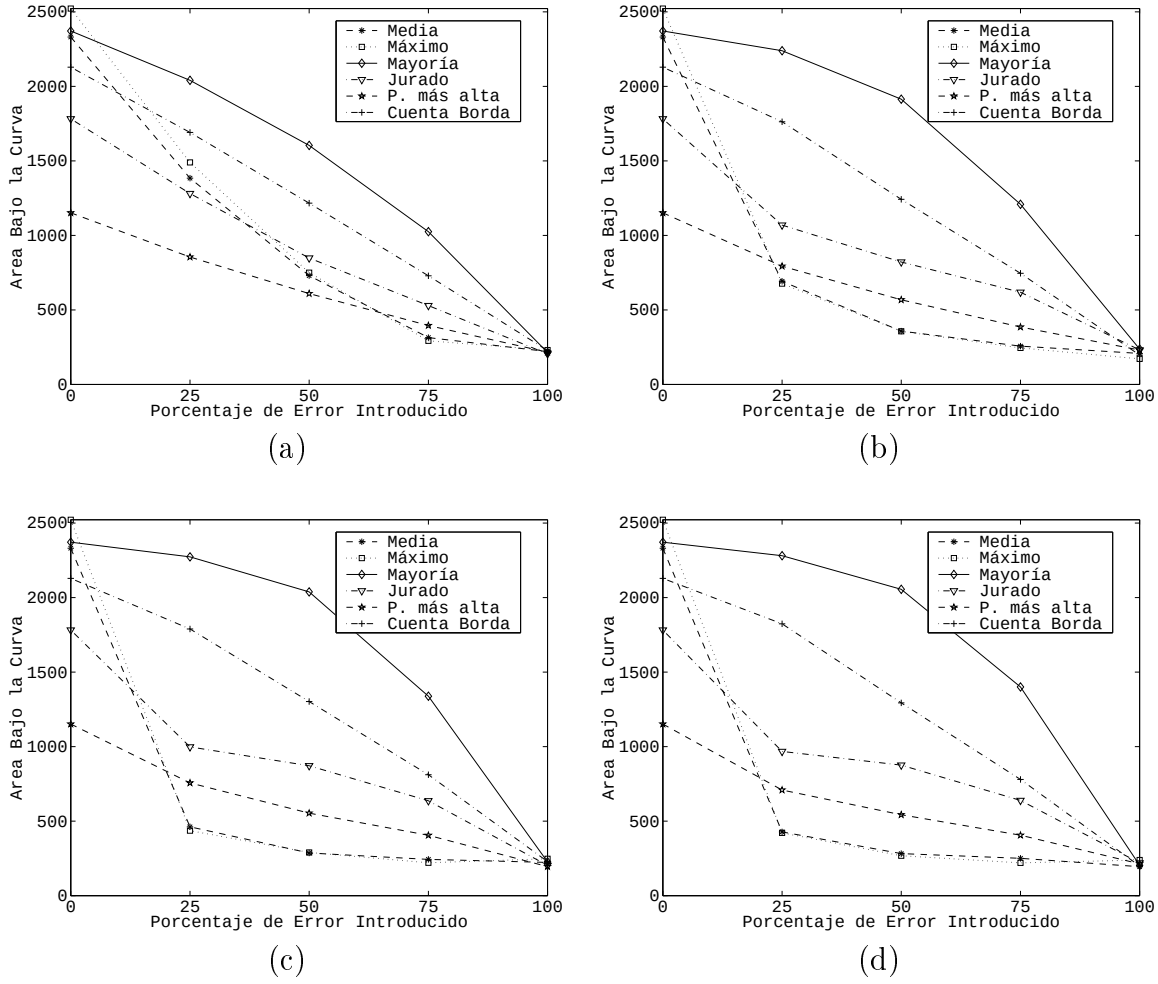

Figura 5. Area bajo la curva de acierto en función de la cantidad de error introducido en los frames de test, utilizando el clasificador VMC, para (a) grado de cercanía (GC) igual al 99 % (b) GC=66 % (c) GC=33 % y (d) GC=0 % .

de persona, habrá que detectarlos a partir de los propios resultados de clasificación [2]. El efecto que se obtiene es equivalente a la introducción de una cierta cantidad adicional de error, como puede verse en la Figura 7, donde se produce un cambio de persona en el frame número 5 de la secuencia. Una vez más se aprecia que la combinación por mayoría de votos es la más robusta, en el sentido de ser la primera en recuperarse de los errores.

[2] Aunque en ciertos casos los niveles inferiores pueden detectar cambios de persona, si se trabaja en un entorno donde las personas pueden ocultarse unas a otras la detección del cambio es más difícil. Cuando una persona oculta a otra la única forma de detectar el cambio es mediante información de profundidad, p.ej. la proporcionada por un medidor láser o por visión estéreo.

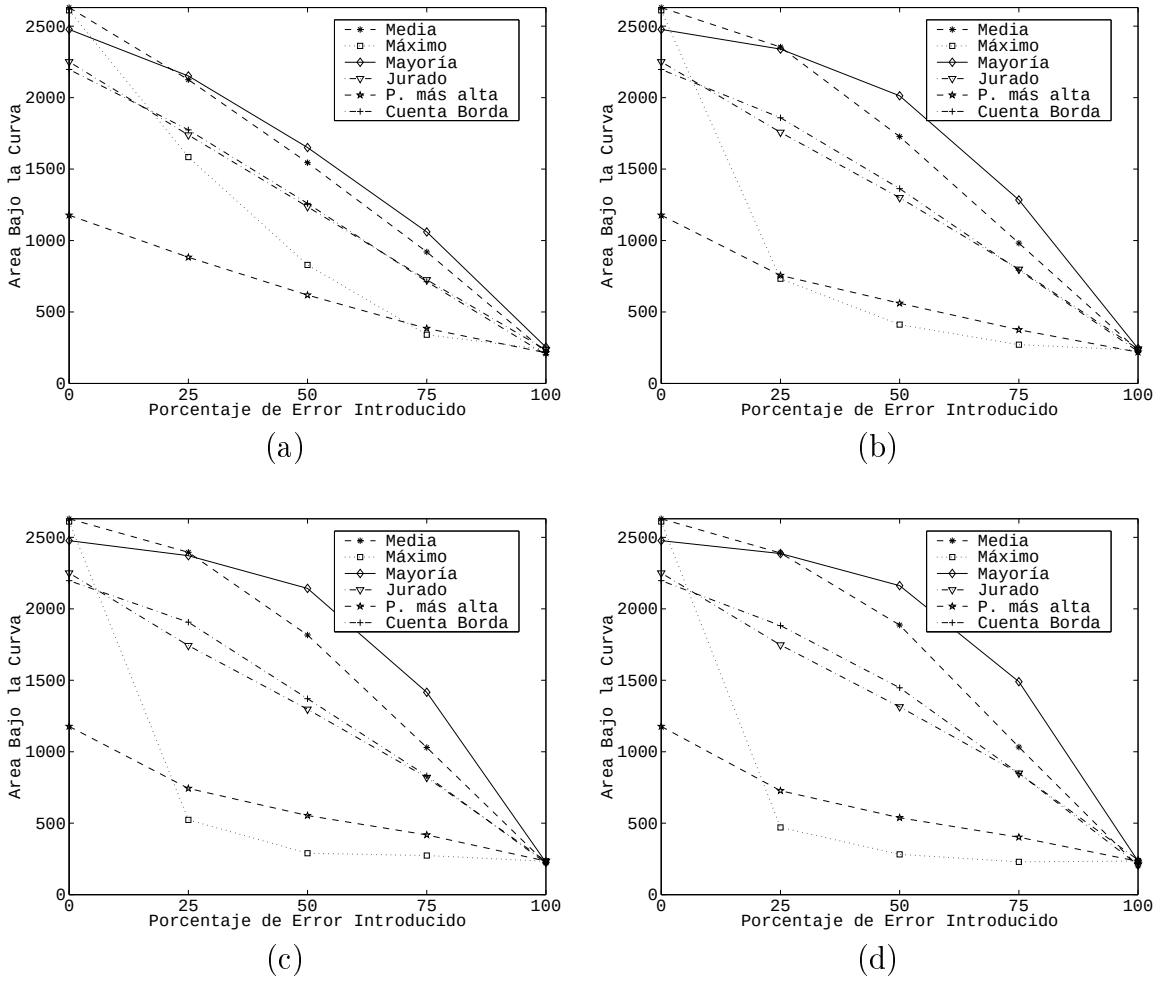

Figura 6. Area bajo la curva de acierto en función de la cantidad de error introducido en los frames de test, utilizando el clasificador SVM, para (a) grado de cercanía (GC) igual al 99 % (b) GC=66 % (c) GC=33 % y (d) GC=0 % .

5 Conclusiones

La posibilidad de mejorar el rendimiento de la clasificación en el problema del reconocimiento de caras mediante la combinación de los resultados de clasificación de varias imágenes de una secuencia es un aspecto importante a tener en cuenta. Se ha estudiado el comportamiento de diversos métodos de combinación de resultados de clasificación. Todos los métodos estudiados son de costo computacional independiente del número de imágenes consideradas. Si bien el máximo y la media son los métodos que producen en condiciones experimentales el mejor comportamiento, se demuestra que tan pronto aumenta el número de falsos positivos introducidos por las etapas de procesado inferiores (detección y normalización), el método más robusto es el de mayoría de votos. El costo computacional de los métodos estudiados es prácticamente equivalente, con lo que se concluye que, para el sistema de reconocimiento estudiado, ejemplo típico de la mayoría de los sistemas

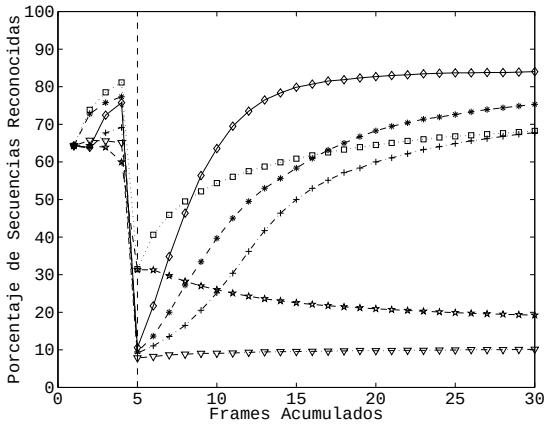

Figura 7. Efecto de una secuencia donde aparecen dos personas, usando el clasificador de media más cercana con distancia euclídea. La leyenda es la misma que la de la Figura 4.

prácticos de reconocimiento de caras, el método más apropiado es el de la combinación por mayoría de votos.

Los métodos estudiados son relativamente simples. Aunque existe cierto paralelismo entre un caso y otro no son directamente aplicables los numerosos métodos existentes de combinación de clasificadores. No obstante, como trabajo adicional se plantea el desarrollo de un método de combinación que sea robusto en condiciones reales de funcionamiento, y que a la vez sea muy rápido en la combinación. El tiempo de combinación es un aspecto muy importante si se quiere diseñar un sistema de utilidad práctica. En situación de funcionamiento real, la combinación debe hacerse lo más rápidamente posible, porque los otros procesos de seguimiento, detección, normalización y clasificación consumirán una gran cantidad de tiempo de proceso. Los seres humanos tardan un tiempo muy pequeño en reconocer caras. Si un sistema de visión artificial pretende interactuar con los humanos no debe demorar mucho su respuesta, o no sería considerado positivamente. Por otro lado, en entornos donde la densidad de personas es alta la detección de cambios ayudaría a aliviar las caídas acentuadas de la probabilidad de acierto. Esta detección podría mejorarse teniendo en cuenta la información que puede proporcionar los niveles inferiores de proceso.

Agradecimientos

El trabajo de O. Déniz está financiado por la beca de formación de personal investigador *D260/54066308-R* de la Universidad de Las Palmas de Gran Canaria. Este trabajo se financió en parte con cargo a los proyectos de investigación DGUI-Gobierno de Canarias *PI1999/153* y *PI2000/042*, y UE/DGES *1FD1997-1580-C02-02*.

Referencias

[1] Yongmin Li, Shaogang Gong, and Heather Liddell. Exploiting the dynamics of faces in spatio-temporal context. In *Procs. The Sixth International Conference on Control, Automation, Robotics and Vision (ICARCV2000)*, Singapore, December 2000.

[2] Osamu Yamaguchi, Kazuhiro Fukui, and Ken ichi Maeda. Face recognition using temporal image sequence. In *IEEE Int'l Conference on Automatic Face and Gesture Recognition*, pages 318–323, Nara, Japan, 1998.

[3] J. Kittler, J. Matas, K. Jonsson, and M.U. Ramos Sánchez. Combining evidence in personal identity verification systems. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[4] Amin Massad, Bärbel Mertsching, and Steffen Schmalz. Combining multiple views and temporal associations for 3-D object recognition. In *Procs. of the ECCV'98*, volume II, pages 699–715, Freiburg, Germany, 1998.

[5] A. Jonathan Howell and Hilary Buxton. Towards unconstrained face recognition from image sequences. In *Procs. of the Second Int. Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, October 1996.

[6] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen. Automatic video-based person authentication using the RBF network. In *First Int. Conf. on Audio and Video-Based Biometric Person Authentication*, Crans-Montana, Switzerland, 1997.

[7] A. W. Senior. Recognizing faces in broadcast video. In *Int'l Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece, September 1999.

[8] S. McKenna and S. Gong. Recognising moving faces. In *Procs. of the NATO ASI on Face Recognition: From Theory to Applications*, Stirling, UK, 1997.

[9] M. Hernández, J. Cabrera, M. Castrillón, and C. Guerra. DESEO: An active vision system for detection, tracking and recognition. In *Procs. of the Second Int. Conf. on Automatic Face and Gesture Recognition*, Killington, Vermont, October 1996.

[10] M. Castrillon, J. Lorenzo, M. Hernandez, and J. Cabrera. Before characterizing faces. In *IX Spanish Symposium on Pattern Recognition and Image Analysis (SNRFAI'01)*, Castellón, Spain, 2001.

[11] An Example of Classifier Combination. Available at `http://www.bangor.ac.uk/~mas00a/demos/pr_example.html`.

[12] T.K. Ho. *A theory of multiple classifier systems and its application to visual word recognition.* PhD thesis, Graduate School of State University of N.Y at Buffalo, May 1992.