Validación Experimental de la Influencia de Oclusiones en Reconocimiento Facial

Ester Gonzalez-Sosa, Ruben Vera-Rodriguez, Julian Fierrez and Javier Ortega-Garcia
Biometric Recognition Group - ATVS, EPS, Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - Campus de Cantoblanco - 28049 Madrid, Spain
{ester.gonzalezs, ruben.vera, julian.fierrez, javier.ortega}@uam.es

***Abstract*—The last research efforts made in the face recognition community have been focusing in improving the robustness of systems under different variability conditions like change of pose, expression, illumination, low resolution and occlusions. Occlusions are also a manner of evading identification, which is commonly used when committing crimes or thefts. In this work we propose an approach based on the fusion of non occluded facial regions that is robust to occlusions in a simple and effective manner. We evaluate the region-based approach in two face recognition systems: Face++ (a commercial software based on convolutional neural networks (CNN)) and an advancement over Local Binary Patterns (LBP) systems considering multiple scales. We report experiments based on the ARFace database and prove the robustness of using only non-occluded facial regions and the limitations of the commercial system when dealing with occlusions.**

I. Introducción

El reconocimiento facial se ha establecido en el campo de reconocimiento biométrico como uno de los rasgos menos intrusivos. Durante la última década, los esfuerzos de investigación han transitado de escenarios controlados y restringidos a los no restringidos y no controlados. La mayoría de estos nuevos esfuerzos se han centrado en la mejora de los sistemas de reconocimiento facial en condiciones de variabilidad tales como la iluminación [1], pose, expresión [2] o imágenes de baja calidad [3]. En los últimos años, el problema de reconocimiento facial bajo oclusiones ha comenzado a recibir atención por parte de la comunidad biométrica [4], [5].

Las oclusiones pueden afectar significativamente al rendimiento de los sistemas de reconocimiento facial. Hay numerosos objetos que al ser puestos, puede ocasionar oclusiones. Las razones para llevarlos también son incontables. Algunas personas pueden cubrir parcialmente su cara indeliberadamente por causas deportivas, estética, tiempo etc.

Por otro lado, los criminales, ladrones, etc. tienden a usar bufandas, gafas de sol o incluso pasamontañas a propósito para evitar ser reconocidos. Esta variedad de oclusiones es muy común en escenarios de video vigilancia y forenses. Un ejemplo de un escenario forense es el caso del atentado del maratón de Boston, donde los dos hermanos implicados [6] llevaban gafas de sol y gorras, haciendo muy difícil su correcta identificación tanto por sistemas automáticos como por expertos forenses. Técnicas del estado del arte basadas en redes neuronales profundas muestran imágenes con oclusiones como errores comunes [7]. Por lo tanto, se requiere aún más investigación para dotar de mayor robustez a sistemas de reconocimiento facial bajo cualquier tipo de oclusiones.

Trabajos anteriores han probado la conveniencia de usar regiones locales en lugar de enfoques holísticos, que tienden a ser más robustos a cambios de pose, iluminación etc. [8]–[10].

El trabajo realizado en [11] analiza el problema del reconocimiento facial bajo oclusiones usando un enfoque basado en parches. Se divide la imagen en 64 bloques y luego los descriptores Local Gabor Binary Pattern Histogram Sequence (LGBPHS) se calculan para cada bloque. El vector resultante es la concatenación de todos los descriptores LGBPHS de cada bloque. La comparación entre dos imágenes se obtiene calculando la distancia de *chi-square* entre los parches no oclusos.

Desde nuestro punto de vista, los enfoques basados en parches no tienen tanto significado como enfoques basados en regiones faciales, ya que las regiones faciales tienen un significado que puede ser interpretado fácilmente por examinadores forenses [10]. Basado en lo anterior, en este trabajo:

- se prueba la robustez de enfoques basados en regiones bajo oclusiones, mostrando que el simple hecho de fusionar regiones faciales no oclusas mejora el rendimiento de enfoques del estado del arte [7].
- se desarrolla un análisis exhaustivo del enfoque holístico y basado en regiones, llevando a cabo experimentos con el sistema comercial Face++ y un sistema LBP multiescala. También se analiza el impacto de las oclusiones en la extracción de regiones faciales y su rendimiento individual, siendo éstas oclusas o no.

Tres escenarios diferentes se consideran: neutral, gafas de sol y bufanda. Los resultados obtenidos prueban los beneficios de descartar las regiones oclusas, logrando una mejora relativa considerable, especialmente para el escenario de gafas de sol.

Este trabajo se estructura como sigue. La Sección II describe la base de datos ARFace usada en este trabajo. La Sección III describe los dos sistemas de reconocimiento facial utilizados. La Sección IV presenta el protocolo experimental seguido y la Sección V presenta los resultados obtenidos. Finalmente, la Sección VI aporta unas breves conclusiones y trabajo futuro.

II. Base de Datos

A diferencia de la gran cantidad de bases de datos existentes que integran múltiples fuentes de variación tales como iluminación, pose o expresión (véase por ejemplo FERET, FRGB, PaSC, MultiPIE), sólo hay dos bases de datos de referencia que integran explícitamente oclusiones: base de datos de ARFace [12] para el dominio 2D y UMB-DB [13] para el dominio 3D. La base de datos ARFace es la que se considera en este trabajo, ya que es la base de datos de referencia más popular en la literatura que contiene oclusiones

reales (hay otras bases de datos que generan las oclusiones de manera artificial).

La base de datos ARFace contiene imágenes de 136 sujetos (76 hombres y 60 mujeres). Cada sujeto está compuesto por 26 imágenes divididas en dos sesiones de 13 imágenes cada una. Las imágenes presentan variaciones con respecto a expresiones (neutra, sonrisa e ira), la iluminación y las oclusiones (gafas de sol y bufanda), y son adquiridas en condiciones controladas. Las sesiones están separadas por dos semanas. Hay algunos sujetos con algunas sesiones incompletas y otros temas que tienen solamente una de las dos sesiones. La base de datos total está compuesta de más de 3300 imágenes. Las imágenes son 576×768 píxeles (anchura $\times$ altura).

III. Descripción del sistema

Como ya se ha mencionado, se consideran dos sistemas de reconocimiento facial diferentes: Face++ (sistema comercial[1] basado en redes neuronales profundas), y sistema basado en LBP (Local Binary Pattern) multiescala.

A. Preprocesado

La etapa de procesamiento previo es diferente de acuerdo con el sistema de reconocimiento facial específico empleado. En el caso de Face++, las imágenes del sistema se dividen en los diferentes parches que alimentan la red neuronal profunda. Para el sistema LBP multiescala, primero se obtienen los puntos de referencia de la cara utilizando un módulo de Face++, se pasa a escala de grises la imagen, y se alinea de acuerdo a la posición de los ojos. Posteriormente se extraen las regiones faciales definiendo regiones rectangulares alrededor de los puntos de referencia.

B. Sistema comercial Face++

Face++ es un sistema de reconocimiento facial comercial, que ha conseguido muy buenos resultados en la conocida competición de reconocimiento facial sobre la base de datos LFW (consiguiendo la segunda posición para el protocolo *sin restricciones* utilizando datos adicionales etiquetados con $0,9950 \pm 0.0036$ de precisión). Face++ se basa en una estructura de red profunda llamada pirámide CNN [14] que puede incorporar de forma natural la compartición de características a través de representaciones faciales multiescala, aumentando la capacidad de discriminación de la representación resultante.

En nuestro caso, la API de la Face++ se utiliza con el fin de obtener resultados de la verificación. Vale la pena señalar que no hay una descripción pública del sistema utilizado por esta API. En primer lugar, las caras que deben compararse son detectadas individualmente utilizando el módulo de detección de la API. Una vez que se detectan las caras, una puntuación de similitud se obtiene llamando a la etapa de comparación. La API de Face++ también da similitud a nivel de cuatro regiones faciales: las cejas, ojos, nariz y boca.

C. Enfoques basados en LBP

1) Sistema LBP multiescala para 4 regiones: Aparte del sistema comercial Face++, se propone en este trabajo un sistema LBP multiescala usando 4 regiones faciales. Una región facial se extrae mediante la definición de una región alrededor de puntos de referencia específicos. Los puntos de referencia son proporcionados por la API de Face++. Tres configuraciones diferentes de puntos de referencia están disponibles: 5, 25 y 83. Después de varios experimentos se observó que el conjunto de 25 puntos era el más robusto frente a oclusiones, y por tanto ha sido el número de puntos seleccionado para poder extraer las regiones faciales (véase la figura 1).

Nuestra implementación del sistema LBP multiescala está inspirado en el sistema propuesto en [15], que describe un rostro a través de características LBP [16] aplicadas a regiones centradas en puntos de referencia y a diferentes escalas. Sin embargo, al seguir un enfoque basado en las regiones faciales, en este trabajo se calcula las características LBP de cada región facial a diferentes escalas en lugar de características LBP de regiones centradas en puntos de referencia. Una de las razones que nos motivan este cambio es que con el enfoque inicial [15], las diferentes versiones escala incluirían regiones oclusas. Con este enfoque somos capaces de aislar las regiones faciales afectadas por oclusiones. Las características de LBP se extraen utilizando el código disponible en [16].

En primer lugar las cuatro regiones faciales diferentes de cara se extraen de la imagen original: las cejas, los ojos, la nariz y la boca (véase la figura 1 derecha). El uso de cuatro regiones se realiza con el fin de hacer comparaciones justas con el software comercial Face++ que también utilizan estas cuatro regiones faciales. Cada región se extrae mediante la definición de una región alrededor de un punto de referencia central. Para calcular el vector de características LBP multiescala para una región y escala específica, primero la región facial se divide en un bloque de 10×10 celdas y, a continuación el histograma LBP se calcula para cada celda. Este procedimiento se realiza para cinco escalas diferentes: 2, 1, $0,5$, $0,25$ y $0,125$. El vector resultante para cada región se logra mediante la concatenación de los 59-vector histogramas LBP de todas las células a diferentes escalas.

Con el fin de reducir la dimensionalidad de las características, una matriz de proyección PCA se estima para cada región utilizando imágenes neutrales, con sonrisa y con ira del conjunto de desarrollo. En todos los casos, el vector de características LBP se reduce a una dimensionalidad de 400 componentes proyectadas. Para la fase de evaluación, las características LBP multiescala asociadas a cada región facial se calculan para luego ser proyectadas en el subespacio PCA. Las medidas de similitud se obtienen utilizando la distancia euclídea. La fusión entre las regiones faciales se realiza a nivel de puntuación utilizando la regla de la suma.

TABLA I
Conjuntos de Desarrollo y Evaluación.

Conjunto	Identidades Hombres	Identidades Mujeres
Desarrollo	1-49	1-39
Evaluación	50-76	40-60

IV. Protocolo Experimental

Toda la base de datos se divide en dos grupos distintos: conjunto de desarrollo y la evaluación de acuerdo con la tabla I. En primer lugar, para estimar la proyección matriz de PCA del sistema LBP multiescala, se utilizan imágenes neutrales,

[1]Para obtener más información, visite http://www.faceplusplus.com/api-overview/). Se utiliza el SDK oficial de MATLAB para Face++ v2.

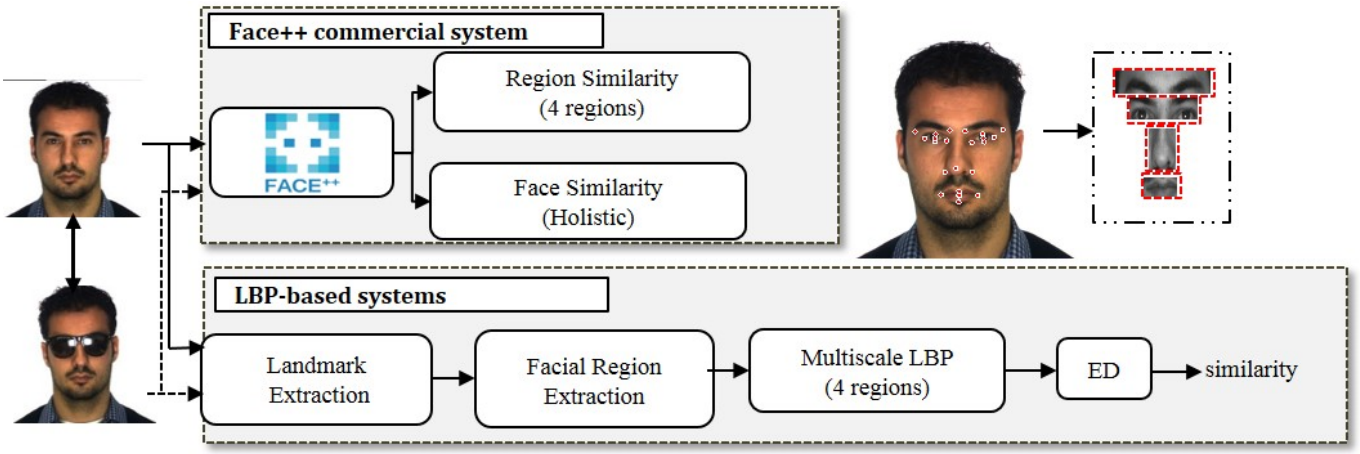

Figura 1. La figura presenta los dos sistemas considerados en este trabajo para hacer frente a las oclusiones en los sistemas de reconocimiento facial. La parte superior del diagrama muestra la utilización del sistema comercial Face++ basado en CNN, dando una similitud global entre dos caras, y también similitudes con respecto a 4 diferentes regiones faciales. La parte inferior presenta el enfoque local considerado: un sistema de múltiples escalas LBP usando 4 regiones faciales. ED se refiere a distancia Euclídea.

con sonrisa e ira imágenes con una iluminación homogénea de las dos sesiones de los sujetos del grupo de desarrollo.

Los resultados se presentan para tres escenarios diferentes: neutral, gafas de sol y bufanda, en términos de EER. En todos los casos, se utilizan las imágenes con una iluminación homogénea de los sujetos del conjunto de evaluación. Para el escenario neutral, las imágenes neutrales de la primera sesión se comparan con las imágenes neutrales de la segunda sesión; para el escenario de las gafas de sol las imágenes neutrales de ambas sesiones se comparan con las imágenes con gafas de sol de las dos sesiones y, por último, para el escenario de bufanda, las imágenes neutrales de ambas sesiones, se comparan con imágenes con bufanda pertenecientes a las dos sesiones. Hasta donde sabemos, no existe un protocolo experimental estándar para esta base de datos. Como el objetivo de este trabajo es estudiar la influencia exclusiva de las oclusiones sobre sistemas de reconocimiento facial, se decide descartar todas las imágenes de iluminación presentes en la base de datos. Vale la pena mencionar que el número de comparaciones del escenario neutral es la mitad con respecto al número de pruebas en los escenarios de gafas de sol y bufanda.

V. Resultados

Los experimentos se llevaron a cabo en la base de datos ARFace siguiendo el protocolo experimental mencionado anteriormente para los dos sistemas: Face++ y sistema LBP multiescala. Para cada sistema y escenario, se muestra el rendimiento de las regiones faciales individuales. El rendimiento de los diferentes esquemas de fusión a nivel de puntuación también se incluyen: i) fusión de todo el conjunto de las regiones faciales; ii) fusión de regiones no oclusas. Para cada escenario específico se define un subconjunto de las regiones no oclusas. En este trabajo, el subconjunto de las regiones no oclusas se ha definido de forma manual para los dos escenarios. Al igual que en [8], el subconjunto de las regiones faciales no oclusas se compone de : nariz y la región de la boca para el escenario de las gafas de sol (2 regiones) y los ojos, las cejas y la nariz para el escenario de la bufanda (3 regiones) .

Los resultados del sistema de Face++ y el sistema LBP multiescala con 4 regiones faciales se muestran en la tabla II. Se puede observar que: i) la fusión de todo el conjunto de las regiones faciales no es siempre mejor que el rendimiento de la mejor región individual, ii) se consiguen los mejores resultados cuando se utiliza la mejor región individual o la fusión de las regiones faciales no oclusas. Al analizar la influencia de las oclusiones sobre el rendimiento de las regiones individuales, se puede ver que las regiones oclusas empeoran seriamente su rendimiento mientras que las regiones no oclusas empeoran ligeramente su rendimiento en comparación con el escenario neutral. El empeoramiento observado en las regiones no oclusas se debe a varias razones. En primer lugar, el rendimiento de la región de la nariz se deteriora en ambos escenarios y sistemas ya que esta región se ve afectada en parte por la presencia de las gafas de sol o bufanda. Por ejemplo, el error de la región de la nariz aumenta de $2,56\%$ de EER a $8,12\%$ de EER para el sistema Face++ y de $12,82\%$ de EER a $27,67\%$ de EER en el sistema LBP multiescala para los escenarios neutral y bufanda. Lo mismo se aplica a la región de la nariz en el escenario de gafas de sol. Además, el rendimiento de las regiones no oclusas también se ve afectada por la pérdida de precisión de los algoritmos automáticos de extracción de puntos de referencias cuando se trata de imágenes con oclusiones.

Al comparar el rendimiento de los dos sistemas, se observa que el software comercial funciona mejor en el escenario neutral, pero el sistema LBP multiescala logra mejores resultados en presencia de cualquiera de las oclusiones consideradas. En concreto, la mejora relativa de la aproximación LBP multiescala con respecto al sistema comercial Face++ es 31% y $7,35\%$ EER para gafas de sol y bufanda, respectivamente.

VI. Conclusiones

En este trabajo se profundiza sobre el problema del reconocimiento de la cara bajo presencia de oclusiones uti-

TABLA II
RENDIMIENTO DE LAS REGIONES INDIVIDUALES Y LOS ESQUEMAS DE FUSIÓN CON SOFTWARE COMERCIAL FACE++ Y EL SISTEMA LBP MULTIESCALA EN TÉRMINOS DE EER%. (*INDICA LAS REGIONES NO OCLUSAS, EL MEJOR RESULTADO DE CADA PAR DE (SISTEMA, ESCENARIO) SE RESALTA EN NEGRITA)

		Individual regions				Fusion	
Sistema	Escenario	Cejas	Ojos	Nariz	Boca	Cuatro regiones faciales	Regiones Faciales No Oclusas
Face++	Neutral	7.69	7.69	2.56	7.69	2.56	**2.56**
Face++	Gafas de Sol	50.71	46.42	33.57*	**12.14***	20	17.14
Face++	Bufanda	15.97*	17.85*	**8.12***	54.46	11.6	11.56
MultiscaleLBP	Neutral	15.38	7.69	12.82	10.25	4.58	**4.58**
MultiscaleLBP	Gafas de Sol	41.46	45.71	22.14*	14.28*	12.85	**11.80**
MultiscaleLBP	Bufanda	16.07*	12.5*	27.67*	49.95	10.71	**10.71**

lizando un enfoque basado en regiones faciales. Evaluamos los sistemas de reconocimiento facial con 4 regiones faciales, logrando mejores resultados que un sistema comercial del estado del arte.

Se han realizado experimentos con dos sistemas diferentes: el sistema comercial Face++ y un sistema LBP multiescala en tres escenarios diferentes: neutros, gafas de sol y bufanda. Se ha demostrado empíricamente la efectividad de la utilización de regiones faciales no oclusas cuando se trata de las oclusiones.

Aunque en este trabajo se presentan resultados para únicamente dos tipos de oclusiones (gafas de sol y bufanda), esta fusión de regiones locales puede ser útil para hacer frente a otros tipos de oclusiones en las condiciones del mundo real. Algunas bases de datos más realistas, tales como Remote Face [17] o LFW [18] serán considerados para el trabajo futuro. La extracción de las regiones faciales en condiciones no controladas será uno de los principales retos a superar.

AGRADECIMIENTOS

CogniMetrics TEC2015-70627-R (MINECO/FEDER) Este trabajo ha sido financiado por el proyecto CogniMetrics TEC2015-70627-R (MINECO/FEDER). Ester Gonzalez-Sosa está financiado con una beca FPI de la Universidad Autónoma de Madrid.

REFERENCIAS

[1] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. on IP*, vol. 19, no. 6, pp. 1635–1650, 2010.

[2] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[3] W. W. Zou, P. C. Yuen, and R. Chellappa, "Low-resolution face tracker robust to illumination variations," *IEEE Trans. on IP*, vol. 22, no. 5, pp. 1726–1739, 2013.

[4] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3d face recognition under expressions, occlusions, and pose variations," *IEEE Trans. on PAMI*, vol. 35, no. 9, pp. 2270–2283, 2013.

[5] X. Tan, S. Chen, Z.-H. Zhou, and J. Liu, "Face recognition under occlusions and variant expressions with partial similarity," *IEEE Trans. on IFS*, vol. 4, no. 2, pp. 217–230, 2009.

[6] J. C. Klontz and A. K. Jain, "A case study on unconstrained facial recognition using the boston marathon bombings suspects," *Tech. Rep Michigan State University*, vol. 119, p. 120, 2013.

[7] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" *arXiv preprint arXiv:1501.04690*, 2015.

[8] K. Bonnen, B. F. Klare, and A. K. Jain, "Component-based representation in automated face recognition," *IEEE Trans, on IFS*, vol. 8, no. 1, pp. 239–253, 2013.

[9] P. Tome, J. Fierrez, R. Vera-Rodriguez, and J. Ortega-Garcia, "Combination of face regions in forensic scenarios," *Journal of Forensic Sciences*, May 2015.

[10] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos, "Identification using face regions: Application and assessment in forensic scenarios," *FSI*, no. 233, pp. 75–83, 2013.

[11] R. Min, A. Hadid, and J.-L. Dugelay, "Efficient detection of occlusion prior to robust face recognition," *The Scientific World Journal*, vol. id 519158, 2014.

[12] A. M. Martinez, "The ARFace database," *CVC Technical Report*, vol. 24, 1998.

[13] A. Colombo, C. Cusano, and R. Schettini, "Umb-db: A database of partially occluded 3d faces," in *IEEE Proc. of ICCV Workshops*, 2011, pp. 2113–2119.

[14] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Proc. CVPR*, 2014, pp. 1891–1898.

[15] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Proc. of CVPR*, 2013, pp. 3025–3032.

[16] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on PAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.

[17] J. Ni and R. Chellappa, "Evaluation of state-of-the-art algorithms for remote face recognition," in *17th IEEE Proc of ICIP*, 2010, pp. 1581–1584.

[18] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. of ECCV Workshop*, 2008.