AVANCES
EN CIENCIAS E INGENIERÍAS
ARTÍCULO/ARTICLE SECCIÓN/SECTION A

Implementing a convolution neural network for recognizing poses in images of faces

Implementación de una red neuronal de convolución para el reconocimiento de poses en imágenes de rostros

Paul Méndez[1], **Julio Ibarra**[1*]

[1] *Universidad San Francisco de Quito. Colegio de Ciencias e Ingenierías - Departamento de Matemáticas. Diego de Robles y Vía Interoceánica, Cumbayá.*
[*] *Autor principal/Corresponding author, e-mail: jibarra@usfq.edu.ec*

Editado por/Edited by: Cesar Zambrano, Ph.D.
Recibido/Received: 14/03/2014. Aceptado/Accepted: 25/08/2014.
Publicado en línea/Published on Web: 19/12/2014. Impreso/Printed: 19/12/2014.

Abstract

Convolutional neural networks belong to a set of techniques grouped under deep learning, a branch of machine learning, which has proven successful in recent years in image and voice recording recognition tasks. This paper explores the use of deep convolutional neural networks in the recognition of horizontal poses outside the plane. We propose a convolutional neural network architecture based on OpenCV open source libraries for classification of images of human faces within seven default poses. We present in details the optimized design of our architecture and our learning strategy.
The classifier trained on a set of 2600 images of sizes: 33×33, 41×41, 65×65 y 81×81, achive an recognition rate of 85 %, higher than the 78 % achieved with the Eigenfaces algorithm, with nearly the same execution time.

Keywords. Convolutional neuronal network, deep neuronal networks, deep learning, machine learning, face recognition, face pose detection.

Resumen

Las redes neuronales de convolución pertenecen a un conjunto de técnicas agrupadas bajo el aprendizaje profundo, una rama del aprendizaje automático que ha probado ser exitosa en los últimos años en tareas de reconocimiento de imágenes y grabaciones de voz. El presente trabajo explora la utilización de las redes neuronales de convolución en el reconocimiento de imágenes de poses horizontales fuera del plano de rostros. Se propone una implementación basada en las bilbiotecas de código abierto OpenCV para la clasificación de imágenes de rostros humanos dentro de 7 poses predeterminadas y se presenta en detalle la arquitectura de la red y la estrategia de aprendizaje.
La implementación entrenada con conjuntos de 2600 imágenes de cuatro tamaños: 33×33, 41×41, 65×65 y 81×81, alcanza una tasa de aciertos promedio del 85 % superior a la obtenida con el algoritmo de Rostros Propios cercana al 78 %, con un tiempo de ejecución similar.

Palabras Clave. Red neuronal de convolución, red neuronal profunda, reconocimiento de rostros, poses, aprendizaje profundo.

Introducción

El procesamiento automático de imágenes para extraer su contenido semántico es una tarea que ha adquirido mucha importancia en los últimos años debido en gran medida al auge de la fotografía digital y sus medios de distribución sobre todo el Internet. Dentro de este contexto, los rostros son especialmente valiosos dado que representan una parte importante de la información semántica contenida en una fotografía.

El reconocimiento facial constituye un área de investigación muy activa en los campos de la Visión Artificial y la Biométrica con aplicaciones en seguridad, robótica, interfaces humano-computadora, cámaras digitales y entretenimiento. Sin embargo, y a pesar del gran esfuerzo dedicado a mejorar los algoritmos de reconocimiento facial, todavía queda mucho por mejorar a fin de que los sistemas puedan producir buenos resultados en tiempo real y bajo condiciones ambientales no controladas. En general el reconocimiento facial sigue siendo un área de activa investigación [1].

En el pasado la mayor parte de las investigaciones rea-

ISSN 1390-5384

lizadas en el área del reconocimiento de rostros se han centrado en el reconocimiento sobre imágenes frontales. Sin embargo, en los últimos años el interés se ha extendido al trabajo sobre imágenes fuera de ambientes controlados, donde el objetivo central es aumentar la robustez de los algoritmos empleados frente a distintas condiciones de resolución, escala, iluminación, expresión facial, pose, entre otros factores [2].

El reconocimiento de pose fuera de plano, tanto horizontal como vertical, resulta importante en sistemas de reconocimiento con sujetos cooperadores, pero más aún para sistemas que buscan el reconocimiento en sujetos no cooperadores, donde es imprescindible desarrollar todo el potencial del reconocimiento de rostros como una técnica biométrica de naturaleza no intrusiva.

El presente trabajo tiene como objetivo explorar la utilización de redes neuronales profundas de convolución para el reconocimiento de poses en imágenes de rostros. Los algoritmos de aprendizaje profundo constituyen una importante alternativa en el area de la inteligencia artificial, que ha presentado resultados muy alentadores en tareas de reconocimiento de imágenes y de audio en los últimos años y que sin embargo se considera, aún está en una etapa temprana de su desarrollo. [3].

En casi todos los usos prácticos de este tipo de algoritmos, la función objetivo es una función altamente no convexa en sus parámetros, con el potencial de presentar muchos mínimos locales. Esto introduce un serio problema ya que no todos los mínimos presentan tasas de error comparables. Como consecuencia en múltiples casos las técnicas usuales basadas en inicialización aleatoria de los parámetros presentan un pobre desempeño [4].

En los últimos años se han presentando diferentes aproximaciones para solucionar este problema. La más importante de estas publicada por Geoffrey Hinton en el 2012 [4] propone pre-entrenar cada capa con un algoritmo de aprendizaje no supervisado, que les permita aprender una transformación lineal de sus entradas que capture las variaciones principales. El pre-entrenamiento es seguido por una etapa final donde la arquitectura se ajusta respecto a un criterio supervisado utilizando optimización basada en el gradiente. Esta estrategia ha reportado mejoras importantes en los algoritmos de aprendizaje de múltiples capas, sin embargo sus mecanismos subyacentes aún son objeto de estudio.

Otro problema presente al entrenar redes neuronales es el "sobre-ajuste". Este es usualmente más critico, si el conjunto de entrenamiento es limitado, dado que los vectores de pesos tienden a usar dependencias entre detectores para ajustarse casi perfectamente al conjunto de entrenamiento, lo cual posteriormente implica un mal desempeño sobre el conjunto de prueba.

Recientemente se ha propuesto como solución el procedimiento conocido como *dropout*, que consiste en omitir aleatoriamente algunos de los detectores de atributos en cada iteración de entrenamiento. De esta forma se reducen las coadaptaciones complejas en las cuales un determinado detector de atributos se vuelve útil solo en el contexto de muchos otros detectores de atributos específicos.

Base de datos	Número de imágenes	Tasa de detección
FERET	2807	99.4 %
FEI	2800	98.6 %

Tabla 1: Tasas de detección del algoritmo en cascada de Viola-Jones.

Sobre la base de datos de imágenes de escritura a mano MNIST,el *dropout* permitió una reducción en la tasa de error para una red de convolución de 5 capas cercana al 19 %. Sobre la base de datos TIMIT para reconocimiento de voz, y una red neuronal de 4 capas conectadas completamente, se consiguió una reducción de la tasa de error sobre un conjunto de prueba de 22.7 % a 19.7 % [5].

En general los autores argumentan que independientemente de la arquitectura el *dropout* permite obtener mejoras moderadas en la tasa de error sobre el conjunto de prueba.

Trabajos posteriores han mejorado esta idea al utilizar el *dropout* como una técnica para promediar modelos y combatir el sobre-ajuste. En este contexto se puede ver el *dropout* en cada actualización como la ejecución de un modelo diferente sobre un subconjunto diferente del conjunto de entrenamiento.

El modelo *maxout* basado en esta idea es una arquitectura de propagación hacia atrás, tal como una red de perceptrón multicapa o una red de convolución, que utiliza una nueva función de activación llamada *unidad maxout*, y se entrena utilizando el modelo *dropout* [6].

Con este método sobre la base de datos MNIST se obtuvo una tasa de error de 0.94 %, que es el mejor resultado a la fecha para algoritmos sin pre-entrenamiento no supervisado. Sobre la base de datos de imágenes CIFAR-10, utilizando una red neuronal de convolución de 3 capas se obtuvo una tasa de error del 12.93 %, que mejora considerablemente la más baja alcanzada para este tipo de algoritmos del 14.05 % [6].

En la siguiente sección se presentarán los experimentos exploratorios diseñados sobre las bases de datos de imágenes de rostros de libre acceso: FERET y FEI y la arquitectura de la red neuronal profunda de convolución utilizada para el reconocimiento de poses. A continuación se presentará las tasas de reconocimiento alcanzadas comparadas contra un algoritmo de Rostros Propios entrenado sobre las mismas imágenes, para después discutir áreas de potencial mejora y desarrollos recientes relacionados con el algoritmo presentado.

Materiales y métodos

Para las pruebas se utilizaron como plataformas de desarrollo una computadora portátil con sistema operativo

	Red Neuronal Profunda de Convolución		Rostros Propios
	Entrenamiento	Prueba	
33×33	0.112 (0.001)	0.157 (0.005)	0.203 (0.005)
41×41	0.098 (0.001)	0.158 (0.003)	0.210 (0.004)
65×65	0.072 (0.001)	0.149 (0.004)	0.210 (0.006)
81×81	0.073 (0.001)	0.154 (0.005)	0.215 (0.006)

Tabla 2: Error promedio (entre paréntesis el error estándar para cada dato), según tamaño de imagen de entrada y algoritmo de reconocimiento.

	Pre-procesamiento (100 imágenes)	RNPC		Rostros Propios	
		Entrenam. (200 epochs)	Prueba (100 imág.)	Entrenam.	Prueba (100 imág.)
33×33	162.8	539.49	1.02	148.93	1.73
41×41	161.5	958.77	1.75	545.4	3.04
65×65	154.8	3105.60	5.27	2206.65	6.71
81×81	162.1	3796.94	8.46	2257.76	9.91

Tabla 3: Tiempos de ejecución en segundos, según tamaño de imagen de entrada y algoritmo de reconocimiento.

Microsoft Windows, procesador Intel I7 de 2.3 Ghz, con 8Gb de memoria RAM y compilador MinGW 4.8 de 32 bits; y una computadora portátil MacBook Pro, con procesador Intel I7 a 2.3 Ghz y 8 Gb de memoria RAM, con compilador CLANG - LLVM 3.3 de 64 bits.

Los rostros utilizados para la etapa de entrenamiento del sistema de estimación de poses se obtuvieron de dos bases de datos de imágenes de rostros de uso público: la base de datos FERET y la base de datos FEI.

El programa (FERET) es administrado por la Agencia (DARPA) (Defense Advanced Research Projects Agency) y (NIST) (National Institute of Standards and Technology). La base de datos consiste en imágenes faciales recogidas entre diciembre de 1993 y agosto de 1996. En 2003 se publicó una versión de alta resolución, 24 bits de color, de estas imágenes. El conjunto de datos incluye 2413 imágenes faciales, representando a 856 persona. Las imágenes a color se encuentran en formato ppm con una resolución de 256×384 pixeles [7].

FEI es una base de datos de libre acceso para investigación y actividades académicas que consiste en 2800 imágenes tomadas entre junio de 2005 y marzo de 2006 en el laboratorio de inteligencia artificial de São Bernardo do Campo, São Paulo, Brasil. Contiene 14 imágenes a color de 640×480 pixeles (10 posiciones e imágenes frontales con diferentes expresiones y condiciones de iluminación) por cada uno de los 200 individuos participantes, para el total de 2800 imágenes. Las imágenes recogen un número igual de mujeres y hombres [8].

Procesamiento previo

Se utilizaron como entrada de la red neuronal de convolución, los rostros detectados y enmarcados por el conocido algoritmo de detección de rostros en cascada desarrollado por Viola y Jones [9].

A fin de construir los conjuntos de prueba y entrenamiento se realizó primero una prueba para determinar las imágenes que podían ser detectadas correctamente por el algoritmo de detección de rostros. Las tasas de detección alcanzadas para las dos bases de datos empleadas se resumen en la Tabla 1.

Las imágenes se normalizaron a imágenes en escala de grises con tamaños estándar de $33 \times 33, 41 \times 41, 65 \times 65$ y 81×81 píxeles cambiando la escala y recortando la imagen según fue necesario para conservar la proporción original del rostro.

A continuación se realizó una ecualización de histograma para mejorar el contraste y el brillo de las imágenes. Esta etapa ayuda a reducir la variación debido a condiciones diferentes de iluminación y es importante para mejorar el desempeño de algoritmos basados en extraer características del rostro [10].

Las imágenes resultantes se convirtieron en arreglos unidimensionales para facilitar la alimentación de los datos a las siguientes etapas.

Finalmente, para preparar los datos de entrenamiento para el sistema de estimación de pose horizontal para cada vista se separaron las imágenes en siete conjuntos según las siguientes categorías: frontal, cuarto de perfil izquierdo y derecho, medio perfil izquierdo y derecho y perfil completo izquierdo y derecho. Las imágenes de perfil corresponden aproximadamente a ángulos de rotación horizontal respecto a la posición frontal de $\pm 22{,}5°$, $\pm 67{,}5°$ y $\pm 90°$ para el cuarto de perfil, medio perfil y perfil completo respectivamente.

Arquitectura de la red neuronal

A diferencia de trabajos con redes neuronales estándar en los que se utiliza un algoritmo adicional para la extracción de las características a ser alimentadas al sistema, la red neuronal profunda de convolución permite delegar la tarea de seleccionar las características importantes a las capas iniciales de convolución de la red. En éstas, la importancia de las características se refleja en los pesos del *kernel* (núcleo) de cada mapa de características.

La red neuronal de convolución utilizada para la estimación de poses se desarrolló a partir de un prototipo de

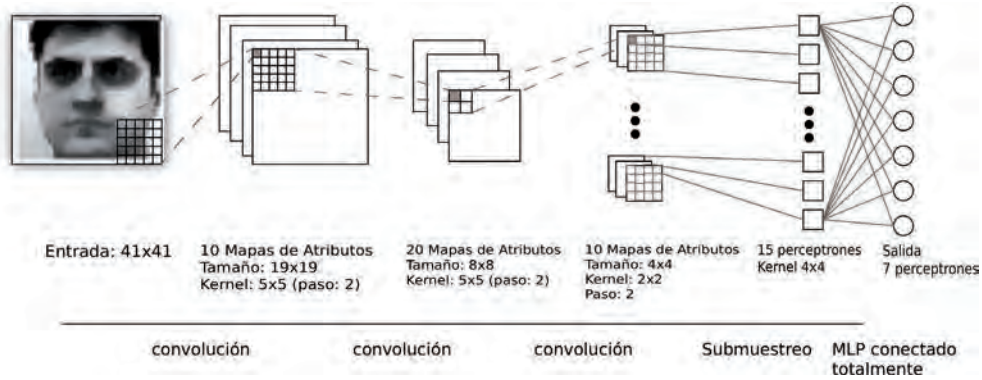

Figura 1: Arquitectura de la red neuronal de convolución usada en la detección de poses.

red neuronal simple con una capa oculta y conexiones completas. Ese prototipo se fue mejorando gradualmente buscando aumentar tanto su rendimiento como su tasa de estimación correcta de poses. El objetivo propuesto para esta etapa preliminar fue alcanzar un error cercano o inferior al 20 %, para después proceder a un ajuste más fino de los parámetros.

La estructura final de seis capas descrita en la sección anterior permitió alcanzar este objetivo sin ningún procedimiento de refinación de sus parámetros, por tanto se aceptó como base para un ajuste más exhaustivo.

Los parámetros iniciales tales como las dimensiones de las imágenes alimentadas a la red neuronal y el preprocesamiento de las mismas se desarrollaron tomando como base trabajos previos, tanto en reconocimiento de imágenes como estimación de poses [11–13]. Al definir estos parámetros se buscó un balance entre la cantidad de información y los requerimientos de memoria y tiempo de procesamiento.

La arquitectura final de la red neuronal está constituida de seis capas: entrada, tres capas de convolución, una capa de submuestreo y una capa de salida tradicional con conexión completa, con función de activación sigmoidal y el número de mapas que se resume en la Figura 1.

La arquitectura de la red neuronal está basada en el trabajo de reconocimiento de dígitos con redes neuronales de convolución desarrollado por LeCun y las sugerencias para optimizar redes neuronales del artículo "Propagación hacia atrás eficiente" del mismo autor [14]. La implementación se basa en el trabajo sobre redes neuronales de convolución para reconocimiento de dígitos de O'Neill [15], cuyo código se encuentra disponible bajo licencia de código abierto MIT X11.

La implementación de la etapa de detección y la red neuronal de convolución se realizó en C++, utilizando la biblioteca para sistemas de visión artificial OpenCV.

OpenCV (Open source computer vision) es una biblioteca de visión artificial de código abierto, originalmente diseñada por Intel en 1999. La librería está escrita en C y C++ y tiene versiones para sistemas Linux, Windows, Mac OS X y algunas plataformas móviles, contiene alrededor de 500 funciones que abarcan diversas áreas de la visión artificial incluyendo el manejo eficiente de imágenes y matrices, reconocimiento de objetos, análisis de imágenes médicas, seguridad, interfaz de usuario, calibración de cámara, visión estéreo y visión robótica [16].

Metodología de entrenamiento

Utilizando las imágenes detectadas correctamente y que por tanto pueden ser sujetas al pre-procesamiento indicado, se elaboraron conjuntos de entrenamiento de aproximadamente 2400 y 2600 imágenes y conjuntos de prueba de aproximadamente 100 y 200 imágenes para la primera etapa de evaluación.

Con el fin de determinar si los parámetros de la red neuronal de convolución no se ajustaron solo a las particularidades de un conjunto de prueba, se buscó la configuración y los parámetros óptimos de la red utilizando solamente la base de datos FERET y posteriormente se utilizaron estos mismos parámetros para entrenar la red con las imágenes de la base de datos FEI y evaluar los resultados sobre un segundo conjunto de prueba.

Para evaluar el impacto del tamaño de imagen alimentado a la red, se entrenó la red neuronal para tamaños de 33, 41, 65 y 81 pixeles, con conjuntos de 2600 imágenes de la base de datos FEI, y se evaluó con conjuntos de 200 imágenes de la misma base de datos. Para cada uno de las 30 evaluaciones realizadas por cada tamaño, los conjuntos de entrenamiento y prueba anteriormente descritos se generaron aleatoriamente.

El algoritmo utilizado para el entrenamiento fue el algoritmo de propagación hacia atrás junto al método estocástico diagonal de Levenberg Marquadt para la optimización de la tasa de aprendizaje η, conjuntamente

Pose	RNPC	Rostros Propios
Perfil completo	0.120 (0.013)	0.153 (0.018)
Medio perfil	0.412 (0.017)	0.547 (0.018)
Cuarto de perfil	0.195 (0.010)	0.273 (0.017)
Frontal	0.037 (0.004)	0.065 (0.005)

Tabla 4: Tasa de error promedio según pose y algoritmo utilizado para imágenes de 65×65 pixeles.

con un término de decaimiento por peso para controlar el sobre-ajuste.

Resultados y Discusión

En el presente trabajo se ha buscado explorar la utilidad de las redes neuronales profundas de convolución como un método para la estimación de poses horizontales fuera de plano, alcanzando como resultado un algoritmo que permite una tasa de estimación de poses cercana al 85 % para las dos bases de datos de prueba estudiadas.

El error promedio alcanzado con la implementación final de la red neuronal profunda de convolución, junto al error obtenido por un algorimto estándar de Rostros Propios se puede ver en la Tabla 2.

Como se puede apreciar en la Tabla 3, a pesar de que el tiempo de entrenamiento es mayor para una red neuronal, no existe diferencias significativas en el tiempo de procesamiento empleado sobre el conjunto de prueba para un algorimto de Rostros Propios (valor-p prueba t pareda menor a 0.01). En general las redes neuronales profundas de convolución requieren de un tiempo sustancialmente mayor para su entrenamiento (los datos presentados en la Tabla 3 no incluyen el tiempo requerido para el ajuste de los parámetros de la red neuronal), pero una vez entrenadas su ejecución sobre datos de prueba es suficientemente eficiente como para su aplicación en tareas de reconocimiento en tiempo real.

La tasa de reconocimiento del 79 % alcanzada por el algoritmo de Rostros Propios en este estudio, es muy similar a la encontrada en estudios similares [17] y ligeramente menor al 85 % alcanzado por la red neuronal. El mismo resultado se confirma al analizar las tasas de error para cada pose.

Como lo revelan los datos de la Tabla 4, en general la tarea de reconocer imágenes en posiciones frontales es sustancialmente más simple que el reconocimiento de otras poses. De entre las poses analizadas aquellas que presetan la mayor tasa de error son las poses intermedias que tienden a ser clasificadas erroneamente entre ellas o con alguna de las poses extremos (frontal y perfil completo).

Un aspecto que se deja para estudios posteriores es el encontrar algoritmos que permitan el escalado y localización uniforme de características del rostro (ojos, nariz, boca) bajo distintos ángulos de rotación. Este tratamiento previo de las imágenes conduce a mejoras en la tasa de reconocimiento de imágenes frontales [18] y puede tener un efecto similar en el reconocimiento de poses. Sin embargo, los algoritmos basados en cascada usualmente utilizados para esta tarea tienen dificultades para detectar la posición de los ojos en rostros con rotaciones horizontales pronunciadas. Bajo estas condiciones usualmente la nariz proyecta cierta sombra sobra los ojos, la cual disminuye considerablemente la precisión de su detección o la impide completamente. Más aún, para imágenes de rostros de perfil completo no es posible detectar ambos ojos y por tanto no es posible usar el procedimiento más común de enmarcado, que emplea la distancia entre estos para escalar las imágenes. Una alternativa para la detección y enmarcado de las imágenes de rostros puede ser el uso de redes neuronales de convolución similares a la arquitectura propuesta por García y Delakis [19] que involucra una red neuronal de convolución de desplazamiento espacial.

Referencias

[1] Zhang, C.; Zhang, Z. 2010. "A survey of recent advances in face detection". *http://research.microsoft.com/apps/pubs/default.aspx?id=132077*, June.

[2] Zhang, X.; Gao, Y. 2009. "Face recognition across pose: A review". *Pattern Recognition*, 42(11):2876–2896.

[3] Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B. 2012. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". *IEEE Signal Process. Mag*, 29(6):82–97.

[4] Hinton, G.; Srivastava, N. 2012. "Improving neural networks by preventing co-adaptation of feature detectors". *arXiv preprint*: 1–18.

[5] Srivastava, N. 2013. "Improving neural networks with dropout". *PhD thesis University of Toronto*.

[6] Goodfellow, I.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. 2013. "Maxout networks". *ICML*.

[7] Phillips, P.; Wechsler, H.; Huang, J.; Rauss, P. 1998. "The FERET database and evaluation procedure for face-recognition algorithms". *Image and Vision Computing*, 16(5):295–306.

[8] Pesquisa, P.; Leonel, L.; Junior, D. 2005. "Relatório Final Captura e Alinhamento de Imagens : Um Banco de Faces Brasileiro". *1-10*.

[9] Viola, P.; Jones, M. 2001. "Rapid object detection using a boosted cascade of simple features". *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR*, 1:I–511–I–518.

[10] Moon, H.; Phillips, P. 2001. "Computational and performance aspects of PCA-based face-recognition algorithms". *Perception-London*.

[11] Le, Q.; Ngiam, J.; Chen, Z. 2010. "Tiled convolutional neural networks". *Advances in Neural*: 1–9.

[12] Vatahska, T.; Bennewitz, M.; Behnke, S. 2007. "Feature-based head pose estimation from images". *7th IEEE-RAS International Conference on Humanoid Robots*: 330–335.

[13] Bouvrie, J. 2006. "Notes on convolutional neural networks". *http://cogprints.org/5869/*.

[14] LeCun, Y.; Bottou, L.; Orr, G.; Müller, K. 1998. "Efficient backprop". *Neural networks*.

[15] O'Neill, M. 2006. "Neural Network for Recognition of Handwritten Digits". *http://www.codeproject.com/Articles/16650/Neural-Network-for-Recognition-of-Handwritten-Digi*.

[16] Bradski, G.; Kaehler, A. 2008. "Learning OpenCV: Computer Vision in C++ with the OpenCV Library". *O'Reilly Media, 1st ed. edition*.

[17] Pang, S.; Kasabov, N. 2006. "Investigating LLE eigenface on pose and face identification". *In Advances in Neural Networks - ISNN 2006, Third International Symposium on Neural Networks, Chengdu, China*: 134–139.

[18] Zhao, W.; Chellappa, R.; Phillips, P.; Rosenfeld, A. 2003. "Face recognition". *ACM Computing Surveys*, 35(4):399–458.

[19] García, C.; Delakis, M. 2004. "Convolutional face finder: A neural architecture for fast and robust face detection". *IEEE Trans. Pattern Anal. Mach. Intell*, 26(11): 1408–1423.