Reconocimiento de Expresiones Faciales

Carim Fadil, Ramiro Andrés Alvarez

Cátedra: Inteligencia Computacional
Docente Tutor: Hugo L. Rufiner
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral,
CC 217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe
{carimfadil,alvarez.ramiroa}@gmail.com
http://www.fich.unl.edu.ar/

Reconocimiento de Expresiones Faciales

Resumen En este trabajo se presenta un método para el reconocimiento de expresiones faciales utilizando la Transformada de Coseno Discreta (DCT) y el Perceptrón Multicapa. Se reconocen las expresiones faciales que corresponden a las siguientes emociones: ira, asco, miedo, alegría, tristeza, sorpresa más la expresión correspondiente a la ausencia de emoción (neutral). Se utiliza la DCT 2D para reducir la dimensión del espacio de la imagen descartando las componentes de altas frecuencias. Los coeficientes no descartados son presentados como entradas al perceptrón multicapa junto con un conjunto de características globales de la imagen. El perceptrón se entrena utilizando el algoritmo de retropropagación estándar. Se encontró que con una arquitectura de una capa oculta con 10 neuronas se obtienen muy buenos resultados de clasificación.

Keywords: expresiones faciales, transformada discreta coseno, perceptron multicapa, redes neuronales, extracción de características.

1. Introducción

A medida que crece el número y la funcionalidad de las máquinas y computadoras con las que tenemos interacción en nuestra vida cotidiana, se vuelve más importante lograr una buena y fluída forma de comunicación con las mismas. Es esperable entonces que pueda haber una interacción menos estructurada entre seres humanos y computadoras/robots. Para que esto sea posible, es necesario poder diseñar sistemas inteligentes que puedan comunicarse con las personas no sólo intercambiando información lógica sino también emocional. Es en este sentido que el reconocimiento de la emoción humana se vuelve un paso muy importante. Mehrabian [1] estableció que en la transmisión de mensajes con carga emocional la parte verbal (palabras) contribuye sólo un 7 % en la construcción del mensaje, mientras que la prosodia (rasgos fónicos que afectan a la métrica de la voz) contribuye un 38 % y las expresiones faciales del hablante lo hacen en un 55 %. Más aún, Ekman [2] demostró que las seis emociones principales se pueden reconocer en las expresiones faciales de manera universal, y que estas expresiones no dependen de la cultura, sino que tienen origen biológico. Estas emociones universales son: ira, asco, miedo, alegría, tristeza y sorpresa. De esta manera, para poder desarrollar sistemas inteligentes que interactúen naturalmente con los seres humanos, se debe implementar el reconocimiento de emociones humanas a partir de las expresiones faciales.

El diseño y uso de sistemas de reconocimiento automático de expresiones faciales presenta distintos problemas:

1. Detección robusta y automática de caras para segmentar la imagen.

2. Procesamiento de imágenes para obtener vectores de características de las expresiones faciales.
3. Representación de los datos para reducir la dimensión del vector de características y mejorar el diseño del clasificador.
4. Diseño e implementación de un clasificador que aprenda los modelos subyacentes de las expresiones faciales.

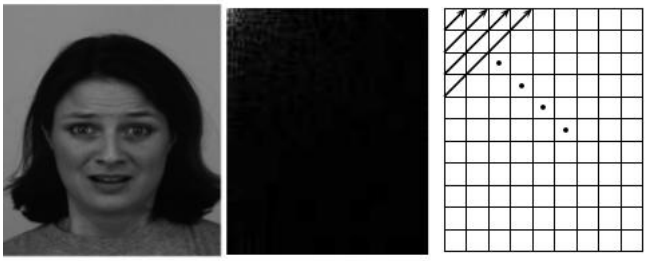
En este trabajo, se utilizan dos bases de datos de expresiones faciales, y se presenta un método que utiliza la Transformada de Coseno Discreta (DCT) en su versión de dos dimensiones, junto con siete parámetros físicos globales de la imagen para extraer características y reducir la dimensión del problema, como en [3]. Estos parámetros físicos son la energía, la entropía, la varianza, la desviación estándar, el contraste, la homogeneidad y la correlación. El clasificador se implementa usando Perceptrón Multicapa (MLP)[4].

2. Extracción de características

Se define una imagen bidimensional de $M * N$ píxeles como una función $f(x, y)$ donde $0 < x < M - 1$ y $0 < y < N - 1$. La DCT 2D de esta imagen está dada por [5]:

$$C(u,v) = \frac{2}{\sqrt{MN}}\alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y). \\ \cos\left[\frac{(2x+1)u\pi}{2M}\right]\cos\left[\frac{(2x+1)v\pi}{2N}\right] \tag{1}$$

para $x = 0, 1, \ldots, M-1$, $y = 0, 1, \ldots, N-1$, $u = 0, 1, \ldots, M-1$, $v = 0, 1, \ldots, N-1$, y $\alpha(w) = \frac{1}{\sqrt{2}}$ para $w = 0$ y $\alpha(w) = 1$ en otro caso.



(a) (b) (c)

Figura 1. (a) Imagen en escala de grises de 762x562 de KDEF, (b) su DCT 2D en escala logarítmica, (c) estrategia para elegir coeficientes.

Como se puede ver en Fig. 1 (b), los coeficientes con valores más altos de la transformada se encuentran en la esquina superior izquierda de la imagen (píxeles claros), que corresponden a las frecuencias espaciales bajas, mientras que el resto de los coeficientes son muy pequeños en magnitud (píxeles casi negros). Para reducir la dimensión de la imagen en Fig. 1 (a) se procede seleccionando sólo los coeficientes DCT de mayor energía como se muestra en la Fig. 1 (c) [6]. De esta manera se está aplicando un filtrado pasa-bajos a la imagen original. Los efectos de reducir la dimensión de la DCT seteando los coeficientes no seleccionados a 0 se pueden ver en la Fig. 2. Con las imágenes de la base de datos KDEF (562x762 píxeles), se seleccionaron 120 componentes de la DCT (0.02 % del total). Esta cantidad fue determinada experimentalmente luego de sucesivas pruebas. El hecho de utilizar este filtrado pasa-bajos no sólo reduce la dimensión de los datos que procesará el MLP, sino que también evita que el clasificador aprenda detalles de cada una de las personas (barba, arrugas, lunares, etc) y contribuye a la capacidad de generalización de la red.

El vector de características se forma entonces con los coeficientes seleccionados de la DCT y con las siguientes características globales de la imagen: energía, entropía, varianza, desviación estándar, contraste, homogeneidad y correlación. Primero se construye una matriz de co-ocurrencia $\mathbf{G}$ con elementos g_{ij}, como se define en [5]. Se define la cantidad

$$p_{ij} = g_{ij}/n$$

donde n es la cantidad de elementos de $\mathbf{G}$. Para mayor explicación sobre la matriz de co-ocurrencia, ver [5]. Luego a partir de esta matriz se extraen las características antes mencionadas. La energía es una medida de uniformidad en el rango $[0, 1]$:

$$\sum_{ij} p_{ij}^2 \tag{2}$$

La entropía es una medida de la aleatoriedad de los valores de la imagen. Se define como:

$$\sum_{ij} p_{ij} \log_2 p_{ij} \tag{3}$$

El contraste es una medida de contraste de intensidad entre un píxel y su vecino sobre toda la imagen. Se define como:

$$\sum_{ij} (i-j)^2 p_{ij} \tag{4}$$

La homogeneidad es una medida de la cercanía espacial de los elementos de la imagen a la diagonal. Se define como:

$$\sum_{ij} \frac{p_{ij}}{1+|i-j|} \tag{5}$$

La correlación es una medida de cuán correlacionado está un píxel con su vecino sobre toda la imagen, y está definida como:

$$\sum_{ij} \frac{(i - m_r)(j - m_c)p_{ij}}{\sigma_r \sigma_c} \tag{6}$$

donde

$$m_r = \sum_i i \sum_j p_{ij}$$
$$m_c = \sum_j j \sum_i p_{ij}$$

y

$$\sigma_r^2 = \sum_i (i - m_r)^2 \sum_j p_{ij}$$
$$\sigma_c^2 = \sum_j (j - m_c)^2 \sum_i p_{ij}$$

3. Diseño del clasificador

Se utilizó un Perceptrón Multicapa[4] con 127 neuronas en la capa de entrada y 7 neuronas en la capa de salida. Las entradas al MLP son los elementos del vector de características descrito en la sección anterior, y la salida es una de las 6 posibles emociones encontradas en la imagen, es decir: ira, asco, miedo, alegría, tristeza, sorpresa; más una salida para el estado neutral (ausencia de emoción). Se utiliza una capa oculta, de acuerdo con el Teorema de Aproximación Universal[4]. Para determinar el número de neuronas en la capa oculta se realizaron sucesivas pruebas; se probó que con 10 neuronas los resultados fueron satisfactorios. Con más neuronas el rendimiento de la red no mejoraba significativamente.

3.1. Normalización de características

La etapa de extracción de características genera un vector con los componentes seleccionados de la DCT y las variables globales mencionadas. Dado que las componentes frecuenciales seleccionadas de la DCT y las variables globales presentan gran variabilidad en órdenes de magnitud, es necesario aplicar una normalización a los datos para obtener una mejor contribución de los mismos en el clasificador. Se mapean los datos al rango $[-1, 1]$ ya que se utilizan funciones de activación sigmoideas simétricas en todas las capas de la red.

Supongamos que $x_1^{(j)}, \ldots, x_n^{(j)}, j = 1, \ldots, p$ son las n muestras del vector de características de las p imágenes de entrenamiento. Se definen entonces los límites superior b_i e inferior a_i como

$$b_i = \beta . \max \left\{ 1, x_i^{(1)}, \ldots, x_i^{(p)} \right\}, i = 1, \ldots, n \tag{7}$$

Figura 2. Imagen original (KDEF) y reconstrucción usando 120 coeficientes.

y

$$a_i = \beta.\,\text{mín}\left\{1, x_i^{(1)}, \ldots, x_i^{(p)}\right\}, i = 1, \ldots, n \tag{8}$$

donde $\beta > 1$ es un factor para extender los límites. Entonces los vectores de entrada $z_1^{(j)}, \ldots, z_n^{(j)}, j = 1, \ldots, p$ a la red están dados por:

$$z_i^{(j)} = 2.\frac{x_i^{(j)} - a_i}{b_i - a_i} - 1, i = 1, \ldots, n \tag{9}$$

Para una imagen desconocida, se utilizarán los factores de normalización obtenidos del conjunto de entrenamiento para obtener el vector de entrada al clasificador[6].

3.2. Entrenamiento

Para el entrenamiento se utilizaron el *Backpropagation Algorithm*[4] (BP) estándar y el algoritmo iRprop descripto en [7] e implementado en la librería FANN (Fast Artificial Neural Network)[1]. Este último es un algoritmo heurístico adaptativo que actualiza los pesos teniendo en cuenta sólo el signo de la derivada parcial del error con respecto a todos los patrones. Para ambos algoritmos se llegaron a resultados similares de clasificación, pero con el último se mejora considerablemente la velocidad de entrenamiento. Para el BP estándar fueron necesarias, aproximadamente, entre 6000 y 10000 épocas de entrenamiento para llegar a los niveles de error deseados. Se probó con distintas tasas de aprendizaje y coeficientes de término de momento, encontrando que con una tasa de aprendizaje $\mu = 0{,}5$ y un término de momento $\eta = 0{,}2$ se lograron las mayores velocidades de entrenamiento. Con el algoritmo iRprop se necesitaron en general entre 1000 y 2000 épocas, aunque estos valores dependen de la conformación de las particiones aleatorias.

[1] http://leenissen.dk/fann/wp/

Se utilizaron funciones de activación sigmoideas simétricas para todas neuronas en todas las capas.

3.3. Validación

Se utilizó validación cruzada[4] con 10 particiones de entrenamiento y prueba. Dado que los datos están balanceados, es decir, se presentan misma cantidad de imágenes de cada una de las emociones en cada base de datos, y misma cantidad de hombres que de mujeres, no fue necesario un balanceo artificial para elegir las particiones. En cada partición, se dividieron los datos en conjuntos de entrenamiento y prueba, asignando un 80 % de los patrones aleatoriamente al conjunto de entrenamiento. Para evitar el sobreentrenamiento y detener el entrenamiento en el pico de máxima generalización, se subdividió el conjunto de entrenamiento en subconjuntos de estimación y validación.

Para cada época, se entrena (estima el modelo) con el conjunto de estimación y se prueba (valida el modelo) con el conjunto de validación. Se detiene el entrenamiento cuando se detecta el pico de generalización, es decir, cuando la tasa de errores en el conjunto de validación comienza a subir o se estanca. Se almacenan los pesos de épocas anteriores para poder utilizar los de mejor rendimiento en el conjunto de validación. Luego, se prueba la red entrenada con el conjunto de prueba para obtener los resultados finales de clasificación.

4. Experimentos y Resultados

4.1. Bases de datos de expresiones faciales

Para este trabajo se trabajó con dos bases de datos de expresiones faciales:

1. The Japanese Female Facial Expression (JAFFE) Database [2]: 213 imágenes de 7 expresiones faciales (las 6 principales y 1 neutral) posadas por 10 modelos femeninas japonesas, y calificadas según 6 emociones por 60 estudiantes japoneses. Se posicionaron las luces para lograr iluminación homogénea en las caras.
2. The Karolinska Directed Emotional Faces (KDEF)[3]: 70 individuos, 35 hombres y 35 mujeres posando 2 veces cada una de las 7 expresiones faciales. Las imágenes fueron tomadas usando una luz suave y homogénea, los individuos usan remeras del mismo color y la posición de la boca y los ojos está centrada en los mismos puntos en todas las fotografías.

En las Figuras 3 y 4 se puede observar muestras de las fotografías usadas en ambas bases de datos.

[2] http://www.kasrl.org/jaffe.html.
[3] http://www.emotionlab.se/resources/kdef

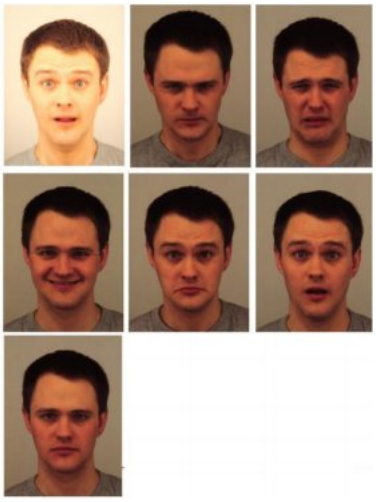


Figura 3. Muestra de Base de Datos KDEF.

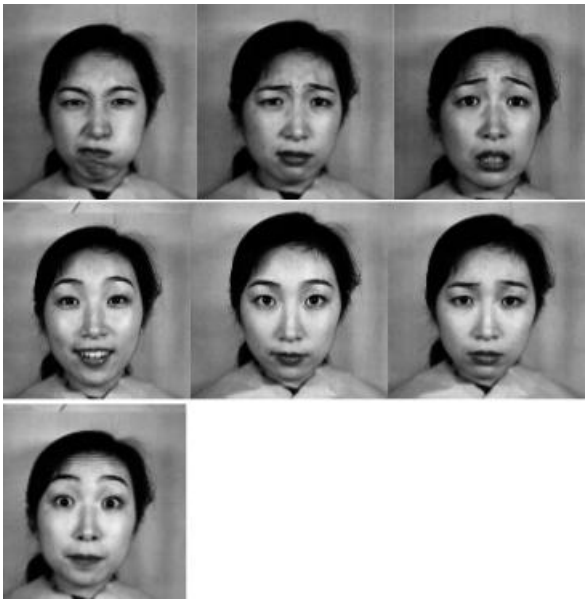


Figura 4. Muestra de Base de Datos JAFFE.

4.2. Resultados

Para medir el desempeño de la red entrenada, se construyeron las matrices de confusión que se pueden ver en las Tablas I y II para las bases de datos KDEF y JAFFE respectivamente.

Estas matrices se generaron de la siguiente manera: luego del entrenamiento, para cada partición de prueba se corre la red sobre cada imagen y se obtiene el valor máximo de salida (salida de la red) que será la emoción que la red detectó en la imagen.

Cada fila de la matriz indica cuántos aciertos y cuántos (y qué tipo) de errores hubo para esa emoción. Se obtuvo un 87 % de acierto para la base de datos JAFEE y un 85 % para KDEF.

De las matrices podemos observar que con la base de datos JAFFE se obtienen mejores resultados para determinadas emociones, y esto se explica en la poca variación que hay en las imágenes de entrada (son sólo 10 modelos y todas mujeres). Sin embargo, se ve que en la base de datos KDEF se obtienen resultados aceptables.

Tabla 1. Matriz de confusión para la base de datos KDEF (Construída sobre todas las particiones de prueba)

	Ira	Asco	Miedo	Alegría	Neutral	Tristeza	Sorpresa
Ira	409	12	0	0	0	19	0
Asco	10	397	15	12	0	5	0
Miedo	23	9	306	0	5	34	22
Alegría	0	15	18	351	0	7	0
Neutral	46	0	14	0	362	8	5
Tristeza	12	8	31	0	46	348	12
Sorpresa	10	0	9	0	10	5	335

Se ve también en las matrices que para ambas bases de datos, las emociones más difíciles de clasificar para el MLP son Miedo (77 % en KDEF y 74 % en JAFFE) y Tristeza (76 % en KDEF y 77 % en JAFFE), dado que el clasificador para ambas bases de datos las confundió entre sí.

En ambas bases de datos, las emociones con las que mejor desempeño se obtuvo fueron Ira (91 % en KDEF y 100. % en JAFFE) y Alegría (90 % en KDEF y 94 % en JAFFE), resultados que también coinciden con lo esperable, ya que son emociones que se reflejan notoriamente en las expresiones faciales.

5. Conclusiones

En este trabajo se logró implementar un método de reconocimiento de emociones a partir de expresiones faciales utilizando Perceptrón Multicapa y la Transformada de Coseno Discreta.

Tabla 2. Matriz de confusión para la base de datos JAFFE (Construída sobre todas las particiones de prueba)

	Ira	Asco	Miedo	Alegría	Neutral	Tristeza	Sorpresa
Ira	98	0	0	0	0	0	0
Asco	0	74	0	0	0	5	0
Miedo	0	18	81	0	0	10	0
Alegría	0	0	0	78	5	0	0
Neutral	0	0	0	0	10	0	0
Tristeza	11	0	0	0	11	73	0
Sorpresa	0	0	10	0	0	0	61

El uso de la Transformada de Coseno Discreta resultó muy adecuado dado que permitió reducir significativamente la cantidad de datos por imagen (0.02 % del total de datos) y al mismo tiempo obtener características de la imagen que distinguen a las expresiones faciales.

El método resultó robusto frente a las dos bases de datos utilizadas: la JAFFE y la KDEF. Se obtuvieron tasas de clasificación del orden del 85 %. Además, la etapa de prueba se resuelve en tiempo real.

Se propone como futuro trabajo la investigación de técnicas alternativas o complementarias a la DCT basadas en procesamiento local de la imagen, de manera de evaluar posibles mejoras en el vector de haicaracterísticas y lograr una mejor representación de los datos.

Referencias

1. Mehrabian, A.: Communication without words. Psychology Today, vol. 2, no. 4, 53–56 (1968)
2. Ekman, P.: Pan-Cultural Elements in Facial Displays of Emotion. Science, vol. 167, 86–88 (1969)
3. Kharat,G. U., Dudul,S. V.: Emtion Recognition from Facial Expression Using Neural Networks. Human System Interaction,2008 Conference On. 422-427 (2008)
4. Haykin, S.: Neural Networks, A Comprehensive Foundation, 2nd Edition. Prentice Hall, New Jersey (1999)
5. González, R. C., Woods, R. E.: Digital Image Processing, 3rd Edition. Prentice Hall (2008)
6. Pan,Z.,Adams, R., Bolouri, H.:Image Recognition Using Discrete Cosine Transforms As Dimensionality Reduction. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing. (2001)
7. Igel, C., Huskel, M.: Grid Improving the Rprop Learning Algorithm. In: Proceedings of the 2nd International Symposyum on Neural Computation. (2000)