Reconocer de Emociones Básicas a Través del Análisis Facial

Ing. M. Jiménez-Vázquez[1], Dr. J.A. Montero-Valverde[2], Dra. M. Martínez-Arroyo[3], MTI. J. Carranza-Gómez[4]

***Resumen*- La expresión facial es una de las formas visuales de comunicación que más utilizan las personas para su interacción con sus semejantes y el mundo que las rodea. En la actualidad se cree que las personas pueden llegar a tener interacción con las computadoras a través de las expresiones, debido a esto, el reconocimiento automático de las emociones humanas a través del análisis facial es una manera natural de interacción en una amplia variedad de aplicaciones. El reconocimiento facial es una etapa importante en este proceso, debido a que de ahí derivan las características utilizadas durante el aprendizaje de los modelos utilizados para la clasificación. En este trabajo se muestran las etapas realizadas en el proceso para reconocer las emociones humanas resaltando la importancia de la detección facial y su posterior clasificación como etapas relevantes del proceso.**

***Palabras clave*- Imagen integral, Boosting, ASEF, HOG, MVS.**

Introducción

Las emociones que los humanos expresan a través del rostro juegan un papel relevante en la vida social. Son señales visualmente observables, conversacionales e interactivas que determinan nuestro foco de atención y regulan nuestra interacción con el entorno y personas vecinas [8]. Asimismo, sabemos que actualmente las computadoras se están convirtiendo en parte de nuestras vidas. Invertimos una cantidad razonable de nuestro tiempo interactuando con dispositivos computacionales de uno u otro tipo (celulares, tabletas, iPhone, videojuegos, etc.). Por el momento, estos dispositivos son, generalmente, indiferentes al estado emocional de las personas. Sin embargo, es del conocimiento común que, para conducir una comunicación humano-humano efectiva debemos tener la habilidad de detectar las señales emocionales de los demás. Por lo tanto, una interacción humano-computadora que no toma en cuenta los estados afectivos de los usuarios pierde una gran parte de información disponible la cual se considera relevante en esta tarea.

Recientemente, la investigación relacionada con los estados afectivos ha sido ampliamente estudiada y existe una creencia creciente de que proveer a las computadoras con la capacidad de entender los estados emocionales de las personas es una tarea importante [1], [2]. Se cree que, con el fin de conseguir progresos en el futuro en las interacciones humano-máquina es necesario que éstas puedan reconocer el estado emocional de los usuarios. Esto se da por entendido debido a la importancia que tienen las emociones en nuestras vidas [3]. La computación afectiva es una rama de investigación que estudia el enlace entre los humanos como entes emocionalmente afectivos y las máquinas como dispositivos con deficiencia emocional [11].

Metodología para el reconocimiento de las expresiones faciales

La metodología utilizada en este trabajo se muestra en la figura 1.1, con el fin de reconocer de manera automática cuatro emociones humanas básicas [9, 10]. Como se observa, la metodología consta de cinco etapas. Una breve descripción de la misma se ofrece a continuación. En la etapa 1 se obtiene el rostro de una persona utilizando la cámara de una computadora, la imagen se toma bajo condiciones ambientales. En la etapa 2 se identifica el rostro de una persona en la imagen tomada con anterioridad, en esta fase se aplica el algoritmo propuesto por Viola y Jones [4]. Una vez que el rostro es detectado en la imagen se procede a la alineación aplicando los Promedios de Filtros Sintéticos Exactos (ASEF) [7], en la etapa 3. Para realizar extracción de las características que representan las diferentes emociones faciales, esta tarea se realiza aplicando la técnica de Histograma de Gradientes Orientados (HOG) [5], esta tarea se lleva a cabo en la etapa 4. El aprendizaje del modelo basado en las Máquinas de Vectores Soporte (MVS) [12], utilizando las características seleccionadas en el paso anterior se realiza en la etapa 5, en esta etapa se tienen que considerar algunas imágenes para el entrenamiento y otras para la evaluación del modelo.

1.-Imagen de Entrada, 2. Detección Facial. Normalización y alineación, 4. Extracción de Características, 5. Clasificación.

[1] Jiménez-Vázquez es alumno de la Maestría en Sistemas Computacionales del I. T. de Acapulco. mario_jv@hotmail.com

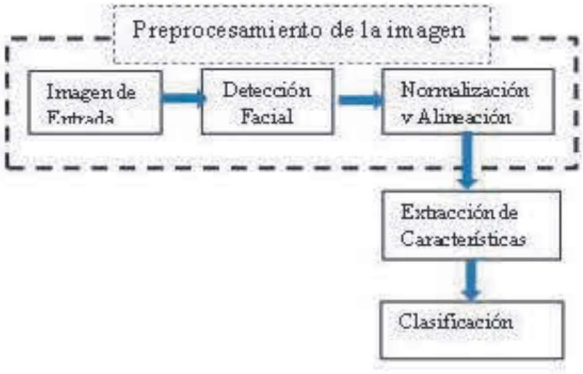

Figura 1. Etapas para el Reconocimiento de expresiones faciales

Detección del Rostro.-La detección del rostro se lleva a cabo aplicando el algoritmo descrito por Viola y Jones [4]. Este algoritmo utiliza una imagen integral para extraer características de forma rápida y precisa, debido a que no trabaja directamente con los valores de intensidad de los pixeles, sino que lo hace a través de una imagen acumulativa que se va formando a través de operaciones básicas que se realizan a medida que se va deslizando en la imagen. La figura 2.a muestra la aplicación de este proceso con el fin de obtener la imagen integral a partir de la lmagen original (Im(x,y)). La imagen integral realiza un desplazamiento de izquierda a derecha y de arriba hacia abajo en la imagen realizando la suma de los pixeles en la localización x,y, a medida que se desplaza (figura 2b), con el fin de obtener la imagen integral aplicando la expresión (1).

En la figura 2.b muestra unos filtros que son utilizados para la extracción de características en una imagen. La suma de los píxeles que se encuentran dentro de los rectángulos blancos se sustraen de la suma de píxeles en los rectángulos grises. Las características de dos rectángulos se muestran en la figura 2b.

a).- b)

Figura 2. Aplicación de la imagen integral y convolución de filtros en el rostro

$$II(X,Y) = \sum_{x'\leq x;y'\leq y} Im(X',Y') \qquad 1$$

Donde:

II(X,Y).-Representa la imagen integral
Im (X',Y').-Representa la imagen original

Condiciones:

x'.-Menor o Igual a x
y'.-Menor o Igual a y

Extracción de características para detectar el rostro

En el reconocimiento facial la extracción de características es aplicada a la imagen utilizando los filtros con bases Haar. Estos filtros son calculados eficientemente sobre la imagen integral y son selectivos en la orientación espacial y la frecuencia, además permiten ser modificados en escala y orientación de acuerdo a las necesidades requeridas, es decir, si se requiere agrandar la imagen se utiliza un múltiplo y si se requiere minimizar se utiliza un divisor en la

escala. En el caso de la detección del rostro, se utilizan los filtros con bases Haar, que a través de la imagen integral recorrerá la imagen facial de izquierda a derecha y de arriba hacia abajo seleccionando la información útil, es decir, la información que contiene los atributos que describen las características del rostro, y desechando la información que no sea de utilidad (los atributos que no contienen características del rostro). Cuando se aplican los filtros Haar, éstos realizan una codificación de diferencia de intensidades en la imagen y no en los pixeles que contiene debido a que éstos trabajan con valores (0,255), generando características de contornos, puntos y líneas, mediante la captura de contraste entre las regiones donde se apliquen los filtros, como se muestra en la figura 4.5, donde se puede observar claramente que se está detectando un rostro en el recuadro que se marca en color verde en la imagen facial obtenida.

Obtención del Rostro

La técnica de *Boosting* fue introducida por Schapire y Freund [6], este es un método de clasificación que utiliza varios clasificadores básicos para formar un único clasificador más complejo y preciso. Los fundamentos se basan en que varios clasificadores sencillos que se desarrollan, cada uno de ellos con una precisión ligeramente superior en una clasificación aleatoria de los ejemplos de entrenamiento, pueden combinarse para formar un clasificador que sea de mejor precisión, siempre y cuando se disponga de un número suficiente de muestras de entrenamiento. La aplicación de clasificadores en cascada ha permitido obtener buenos resultados en las muestras de entrenamiento, entre mayor sea el número de muestras, habrá mayor precisión en los resultados obtenidos, como se muestra en los trabajos realizados por Viola y Jones [4].

Para aplicar la técnica de *Boosting* primero se debe establecer un algoritmo de aprendizaje sencillo (clasificador débil o base), que será llamado repetidas veces para crear diversos clasificadores base. Para el entrenamiento de los clasificadores base se emplea en cada iteración, un subconjunto diferente de muestras de entrenamiento y una distribución de pesos diferente sobre las muestras [6]. Entre mayor sea el número de muestras de entrenamiento mayor será la precisión en la clasificación de las características. Finalmente, estos clasificadores base se combinan en un único clasificador que es mucho más preciso que cualquiera de los clasificadores base por separado. Como resultado de la combinación de los filtros Haar, la técnica de Boosting y el algoritmo de Viola y Jones, se muestra la imagen de la figura 4.5.

Localización de los Ojos

Generalmente los algoritmos de búsqueda de ojos utilizan las coordenadas (x, y) para ubicar los píxeles del centro de los ojos izquierdo y derecho en las imágenes frontales. Para que esto sea verdadero, el algoritmo debe devolver la ubicación del ojo proporcionando cierta tolerancia, medida típicamente como una fracción de la distancia interocular, es decir, la distancia entre los centros de los ojos en el rostro. Específicamente en este trabajo, se propone la utilización de la clase de filtros denominada Promedio de Filtros Sintéticos Exactos (ASEF) [7], por dos razones importantes en el desarrollo de este trabajo. En primer lugar, se especifica una superficie de respuesta de correlación completa para cada instancia de entrenamiento durante la construcción del filtro. En segundo lugar, el resultado de los filtros utilizados en cada imagen de entrenamiento se promedia para mostrar el objeto.

Delimitación de la Imagen

Después de haber detectado el rostro en la imagen de entrada, el siguiente paso es delimitar el contorno facial de la imagen (rostro) para observar que están presentes todos y cada uno de los componentes faciales (ojos, boca, nariz, cejas, frente). Esto se realiza con la finalidad de dejar solamente el rostro de la imagen que se muestra en la figura 4, dejando a un lado los otros elementos componentes del rostro como: orejas, pelo, cuero cabelludo y cualquier otro objeto (aretes o algún tatuaje), que pueda estar presente en el rostro y que puedan alterar el contenido de la imagen facial. Este punto es muy importante debido a que la imagen debe estar completamente despejada de cualquier objeto que pueda proporcionar información inadecuada que interfiera en el siguiente proceso. El rostro delimitado y alineado durante la fase de preprocesamiento se muestra en la figura 3.

Figura 3. Imagen original y detección del rostro aplicando los filtros de correlación sintética.

La imagen de la figura 3 muestra en el recuadro de color verde, el rostro detectado en la imagen de entrada tomando como puntos de referencia los ojos para una mejor ubicación, con la aplicación de los promedios de filtros sintética exactos de correlación.

Descriptor de Características

Un descriptor de características es una representación de una imagen que la simplifica al extraer información útil y descartar información irrelevante. Típicamente, un descriptor de características convierte una imagen 3D a un vector (conjunto de características de longitud **n)**, en el descriptor de características HOG.
En el descriptor de características de HOG, la distribución (histogramas) de las direcciones de los gradientes (gradientes orientados) se utilizan como características. Los gradientes (derivados **x** e **y**) de una imagen son útiles porque la magnitud de los gradientes es grande alrededor de los bordes y esquinas (regiones de cambios abruptos de intensidad) y sabemos que los bordes y esquinas contienen mucha más información sobre la forma del objeto que las regiones planas del mismo. En una investigación realizada por Dalal y Triggs [5], encontraron que se pueden definir vectores de características de baja dimensión que son sensibles al contraste. En dichos estudios se ha encontrado que los rendimientos en algunas categorías de objetos mejoran el uso de características sensibles al contraste, mientras que algunas categorías se benefician del contraste de características insensibles. Por lo tanto, en la práctica, se utilizan vectores de características que incluyen ambos contrastes sensitivos y no sensitivos.

Como se ha mencionado anteriormente, en la práctica se puede utilizar una proyección analítica utilizando vectores dimensionales. En este trabajo de investigación se utilizan 108 vectores dimensionales, definidos por 27 sumas sobre diferentes normalizaciones, uno para cada canal de orientación, **(9 insensibles al contraste y 18 sensibles al contraste) y 4 dimensiones que capturan la energía del gradiente general en bloques de diez celdas (i, j). Por lo tanto, el mapa de características final tiene un vector de 31 dimensiones [7].**

El vector final de características se calcula con los siguientes datos:

W = ancho de la imagen entre el tamaño de la celda = 80 / 8 = 10
H = alto de la imagen entre el tamaño de la celda = 96 / 8 = 12
HOG = dimensión del vector de características de HOG = 31

Por lo tanto, la fórmula utilizada es la siguiente:

Tamaño final del vector de características = W x H x HOG

Tamaño final del vector de características = 10 x 12 x 31 = 3720.

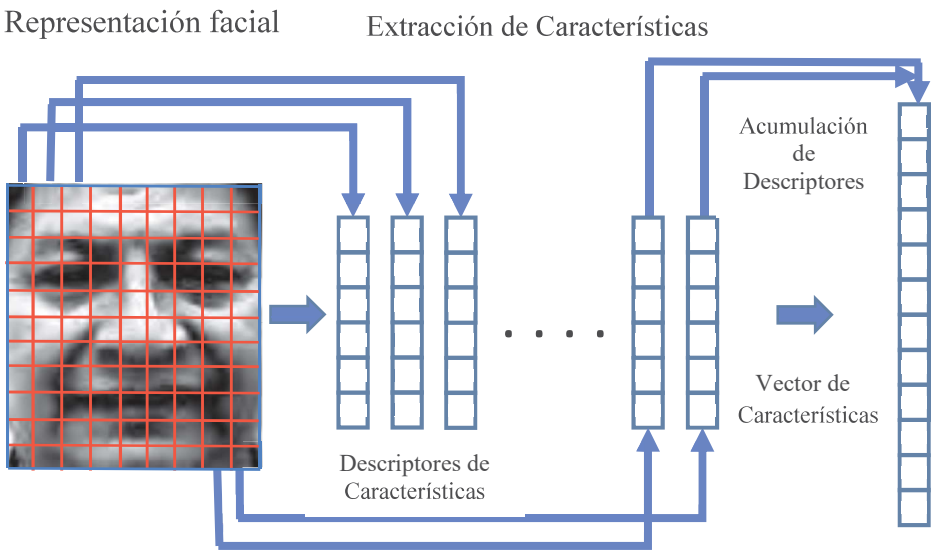

Figura 4. Rostro dividido en celdas para obtener los descriptores de características que se forman utilizando HOG, acumulados para formar el vector de características final.

Máquina de Vectores de Soporte

Una Máquina de Soporte Vectorial (MVS), es un sistema de aprendizaje automático que se ha utilizado para resolver problemas de clasificación y regresión de manera muy eficiente [12], estas máquinas son capaces de clasificar muestras en dos posibles conjuntos de información "positiva" y "negativa", que en este caso cuando realiza la detección de una imagen facial corresponden a "rostro" y "no rostro". Desechando la información que no resulta útil para la detección del rostro (información negativa) respectivamente. Para realizar este proceso, se requiere de un entrenamiento previo de la máquina con esta información, por lo que se le introduce ejemplos de información "positiva" y "negativa", que corresponden a las imágenes faciales para que realice dicha clasificación. A cada uno de los vectores finales que se formaron con la suma de todos los descriptores de características de los bloques de celdas que componen la imagen se les asigna un número de etiqueta (correspondiente a cada imagen) para llevar a cabo su identificación y clasificación para que posteriormente alimenten el algoritmo de aprendizaje. El proceso de alimentación de la Máquina de Vectores de Soporte con los parámetros (valores numéricos) que representan a cada uno de los vectores finales de características que describen los estados emocionales de las imágenes almacenadas en la base de datos, se realiza para clasificar a éstas con la finalidad de reconocer el estado emocional detectado en ella.

Resumen de resultados

Tabla 1. Muestra de los resultados que se obtuvieron al clasificar cuatro estados emocionales básicos aplicando esta técnica de normalización. En este caso la confiabilidad del clasificador es del 92% mostrando un error del 8%, el incremento de la confiabilidad se debe a que las características que interesan para formar el vector utilizado en las etapas de entrenamiento y evaluación fueron resaltadas en la normalización.

	Contento	Enojado	Neutral	Sorpresa
Contento	221	3	1	13
Enojado	3	217	12	6
Neutral	5	9	223	1
Sorpresa	12	1	7	218

Conclusiones

En este trabajo se mostró una arquitectura basada en histogramas de gradientes orientados y máquinas de vectores soporte para realizar esta tarea. Aunque estas técnicas ya se han utilizado de forma conjunta en otras aplicaciones (reconocimiento de personas, seguimiento de objetos) no se habían enfocado al reconocimiento del estado emocional humano. En este trabajo se utilizaron un total de 952 imágenes, es decir, participaron 238 personas, la mayoría estudiantes del Instituto Tecnológico de Acapulco, a cada uno de ellos se les pidió que actuaran de forma natural al mostrar sus emociones. Asimismo, las imágenes fueron capturadas en condiciones ambientales normales. Por lo tanto, se concluye lo siguiente.

1.- La arquitectura planteada inicialmente, basada en utilizar la técnica de HOG para la extracción de características integrada con las MVS´s como mecanismo de clasificación genera resultados satisfactorios cuando se trabaja con imágenes obtenidas bajo condiciones no controladas.

2.- Los clasificadores basados en MVS´s realmente obtienen resultados satisfactorios cuando trabajan con cantidades limitadas de datos.

Comentarios finales

Es conveniente integrar un mayor número de emociones para que esta herramienta tenga un mayor número de aplicaciones. El objetivo de esta herramienta es integrarla en un sistema tutor inteligente con el fin de que ofrezca una tutoría cuando el estudiante se encuentre en condiciones de aprender. Sin embargo, con el fin de no limitar su utilidad es necesario integrar más emociones.

Referencias

[1] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas Huang. Human computing and machine understanding of human behavior: A survey. In ACM International Conference on Multimodal Interfaces, pages 239 – 248, 2006.
[2] Peter Robinson and Rana el Kaliouby. Computation of emotions in man and machines. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535):3441–3447, 2009.
[3] Jeffrey F. Cohn. Foundations of human computing: Facial expression and emotion. In ACM International Conference on Multimodal Interfaces, pages 233–238, 2006.
[4] P. Viola and M.J. Jones, "Robust real-time object detection," *Int. Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154, Dec. 2004.
[5] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection".
Book: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.
[6] SCHAPIRE, R and FREUND, Y. A decision theoretic generalization of on-line learning and application to boosting. AT&T Bell Laboratories. USA, 1995.
[7] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan "Detección de objetos con entrenamiento de modelos discriminatorio basados en partes".
[8] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *IEEE Trans. Image Processing,* vol. 16, no. 1, pp. 172-187, 2007.
[9] P. Ekman and W.V. Friesen, "Constants across cultures in the face and emotions," *J. Personality Social Psychology,* vol. 17, no. 2, pp. 124-129, 1971.
[10] M. Pantic, I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," *Proc. IEEE conf. Systems, Man and Cybernetics,* vol. 4, pp. 3358-3363, Oct. 2005.
[11] [29] Rosalind W. Picard, Affective Computing. Cambridge (MA): MIT Press. 1997.
[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.