Las tres etapas principales son: detectar el rostro, extraer los atributos y clasificarlos. Esta fase del estudio parte del supuesto de contar con una entrada de datos donde se enfoca el rostro, ya sea para una imagen fija o en movimiento.

Sistema de Codificación de Acciones Faciales (FACS) (Pantic 2006). Este consiste en un profundo estudio de los músculos faciales y una representación de las contracciones de estos en un conjunto de valores llamado Unidades de Acción (AU). Una unidad de acción es, por ejemplo, el levantamiento de la parte interior de la ceja (AU1), y otra el levantamiento de la mejilla (AU6). Las unidades no necesariamente tienen una equivalencia con un músculo, para una unidad de acción pueden intervenir varios músculos faciales, o un mismo músculo determinar por sí solo más de una unidad.

K. Khorasani usó la imagen del rostro con la llamada clasificación “neutral” para comparar los rasgos extraídos de esta contra los de la que se quiere clasificar (Khorasani, 2004), esta variante por supuesto implica el obvio inconveniente de la necesidad de contar de antemano con la imagen del rostro “neutral” y su procesamiento.

Otros trabajos (Azcarate, 2005) cuentan con el inconveniente de tener que fijar a mano marcadores en la imagen para que el sistema le pueda seguir el rastro a los movimientos, los cuales son además analizados en una secuencia de imágenes. Con limitantes parecidas, Cohen (et. al, 2003) trabajan con una entrada continua de video la cual es preparada en la imagen inicial, marcando puntos como los bordes de los ojos y la boca, y luego hacen coincidir la imagen del rostro neutral con estas marcas para construir una malla 3D. Con la malla extraen los movimientos de estructuras e identifican los rasgos y las dependencias entre estos. Utilizan marcas temporales para ir procesando el desarrollo de una expresión hasta clasificarla en el clímax de su proyección.

Otros trabajos (Azcarate, 2005) cuentan con el inconveniente de tener que fijar a mano marcadores en la imagen para que el sistema le pueda seguir el rastro a los movimientos, los cuales son además analizados en una secuencia de imágenes. Con limitantes parecidas, Cohen (et. al, 2003) trabajan con una entrada continua de video la cual es preparada en la imagen inicial, marcando puntos como los bordes de los ojos y la boca, y luego hacen coincidir la imagen del rostro neutral con estas marcas para construir una malla 3D. Con la malla extraen los movimientos de estructuras e identifican los rasgos y las dependencias entre estos. Utilizan marcas temporales para ir procesando el desarrollo de una expresión hasta clasificarla en el clímax de su proyección.

M. Karthigayan y M. Rizon (2008) trabajaron los ojos y la boca aproximados con construcciones de elipses. Para esto, utilizaron un algoritmo genético con el objetivo de encontrar las aristas de las elipses que mejor aproximaban las imágenes, usando una elipse regular para los ojos y una elipse irregular o compuesta (de 2 aristas menores) para la boca, por la diferencia entre el labio superior y el inferior. Además aplicaron una técnica de contorno sobre una escala negro y blanco para la definición de las estructuras y encontraron con pruebas calculadas a mano de que las clases que querían obtener presentaban diferentes valores de las aristas de las elipses, por lo que este conjunto serviría de entrada al problema de clasificación.

R. Lienhart y J. Maydt (2002) que propone la utilización de las llamadas características tipo Haar para la detección de objetos en una imagen digital, inspiradas a su vez en las funciones ortonormales de la secuencia Haar propuesta por el matemático Alfred Haar en el 1909.

Otro algoritmo muy utilizado en esta rama es la transformada de coseno discreta (DCT por sus siglas en inglés). DCT es un algoritmo altamente explotado en el área de compresión de imágenes. La más usada es la DCT-II bidimensional por sus propiedades de suavidad en los extremos de los intervalos.

El principio de compresión detrás de las DCT está basado en las series de Fourier. Calculando los coeficientes para los distintos vectores de la base, se logra la aproximación.

L. Ma y K. Khorasani (2004), quienes usaron los coeficientes más significativos de la DCT como la entrada de su clasificador, que era una red neuronal. Ellos además hicieron cierto pre procesamiento, restándole una imagen neutral a la que se iba a clasificar, además de tener todas estas ya normalizadas con dimensiones y características específicas. Después de aplicar la DCT, hicieron varias pruebas con distintas dimensiones de la matriz de coeficientes significativos a seleccionar, de los obtenidos con la DCT, para quedarse con la óptima y utilizar esta matriz como su conjunto de patrones que determinarían la clasificación.

Esta variante se ha usado frecuentemente en trabajos de reconocimiento de expresiones en imágenes rostros (Pan, 2000), pues constituye un procedimiento computacionalmente eficiente de trabajar con una representación de la imagen

M. N. Dailey y G. W. Cottrell (2002), donde utilizaron los llamados filtros Gabor como patrones de la imagen para la entrada del clasificador. El principio de los filtros Gabor es compartido con la DCT en los trabajos de Fourier. Estos son también funciones con componentes trigonométricos para los que se definen diferentes orientaciones angulares y de conjunto constituyen una representación optimizada de la imagen.

Una de las técnicas de extracción de rasgos más comunes en reconocimiento de emociones faciales son la extracción basada en píxeles (Subramanian et al., 2012), los filtros Gabor, las transformadas de Curvelet y los patrones binarios locales

M. Pantic (2000) Divide los enfoques de extraer rasgos en tres formas: tomando el rostro completo (holístico), tomando el rostro como un conjunto de componentes como ojos, boca, cejas, etc. (analítico), o una combinación de estos dos enfoques (hibrido).

En enfoques analíticos se presentan modelaciones del rostro basándose en puntos distribuidos en el rostro, con cierta similitud a los puntos de las Unidades de Acción de Ekman. En los holísticos se ponen de ejemplo las mallas 3D y modelos espacio-temporales de movimientos (para secuencias de imágenes). Para los híbridos plantean que normalmente se utilizan puntos para determinar la posición inicial de alguna plantilla.

Un sistema ideal debe ser capaz de hacer todas las etapas automáticamente (sin que medie ningún humano), saber manejar condiciones de la foto como iluminación, orientación e inclinación del rostro, variación de tamaño, ruido y desenfoque en la calidad de la imagen, obstáculos en el rostro como el pelo y espejuelos, distinguir todas las expresiones y detectar las 44 AU de FACS, tener aprendizaje adaptativo, operar en tiempo real y asignar etiquetas de interpretación cuantificadas y múltiples. (Pantic, 2000)

El uso de un modelo derivado del sistema FACS y de las Unidades de Acción de Ekman ha sido lo más frecuente y ha mostrado altos niveles de acierto en la clasificación. Combinar esto con técnicas de regiones de interés también ha mostrado resultados elevados.

Como clasificadores se utilizaron con éxito las redes neuronales, algoritmos bayesianos, kNN y máquinas de soporte vectorial (SVM), así como algoritmos compuestos y derivados de estos.

Las necesidades para cada sistema varían, pero la base de datos de Cohn-Kanade ha sido la base más utilizada y recomendada por los trabajos en el área debido al volumen de sujetos y confiabilidad de la clasificación, que cuenta con codificación de las Unidades de Acción, cuenta con secuencias de imágenes partiendo de la neutra que permiten el trabajo tanto para sistemas de imágenes fijas como para imágenes en progresión, y tiene un balance adecuado en los sujetos que la componen en términos de sexo, etnia y edad. Aún así, en la experiencia de los autores de este trabajo, se detectaron varias inconsistencias en esta base de datos.

<http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992015000100005>