Capítulo 4 Implementación del sistema

4.1 Hipótesis

A continuación se detalla la implementación de las diferentes fases del método para detección de gestos. La idea principal del funcionamiento de este método como segmentación y filtro de ruido se debe a que el movimiento más cercano a la cámara es el mayor. Esto se vuelve visible al calcular las diferencias en referencia con el cambio más alejado percibido por la lente. Si el cambio más grande es el de la mano haciendo el gesto frente a la cámara, será entonces el cambio más grande percibido por la cámara (HOO, SIANG, DUNG, YU, & CHOON, 2005).

Podríamos mover ciertas partes del cuerpo de manera inconsciente pero al levantar la mano y apuntar hacia la cámara, el movimiento más grande será el del gesto. Al tomar en consideración lo anterior, se puede decir que por la suposición este es un método ingenuo, pero podrían añadirse filtros posteriormente para mejorar la efectividad del método. Partiendo de esta suposición, para detectar y rastrear el movimiento, se utiliza el cruce de los histogramas generados a partir de la binarización de la diferencia de las imágenes. A pesar de que el tamaño o la velocidad de los cambios en la imagen podrían también utilizarse para filtrar el ruido detectado no deseado, o montar un sistema de visión estéreo, implementarlo no se encuentra en el alcance de este proyecto. A continuación el detalle del método, sus fases y sus variaciones.

4.2 Reconocimiento facial

El reconocimiento facial fue realizado utilizando una librería con el método de detección robusta de rostros de Viola-Jones (Viola & Jones, 2004) y utilizando las características tipo Haar ya entrenadas de la librería de Intel OpenCV (OpenCV, 2015). Este método se basa principalmente en los siguientes componentes;

- Descriptores tipo Haar.
- Imagen integral.
- Adaboost
- Análisis en cascada

A continuación se hará una breve descripción de los diferentes pasos que componen el método de detección rápida de Viola-Jones.

4.2.1 Descriptores tipo Haar.

Un detector de objetos basado en descriptores tipo Haar toma principalmente tres tipos de filtros digitales, para detectar orillas, líneas y diagonales. El valor de una característica de dos rectángulos es la diferencia entre la suma de los pixeles dentro de las dos regiones. Estas regiones tienen el mismo tamaño y la misma forma y son adyacentes vertical u horizontalmente. Una característica de tres rectángulos calcula la suma dentro de los rectángulos de las orillas y lo resta de la suma del rectángulo del centro. Finalmente una característica basada en cuatro rectángulos calcula la diferencia entre los pares diagonales de los rectángulos (ver figura 22).

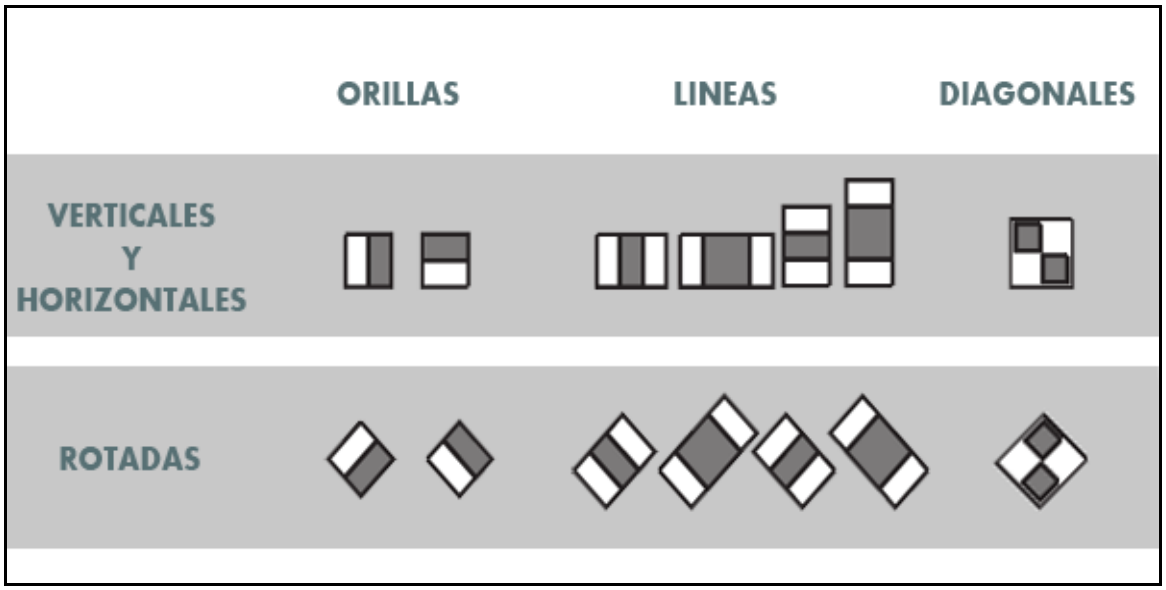

FIGURA 22 CARACTERÍSTICAS DE HAAR

4.2.2 Imagen integral.

Los detectores de la cascada trabajan sobre una imagen con los valores en escala de grises llamada imagen integral o integral de la imagen donde cada pixel contiene la sumatoria de todos los valores acumulados de sus intensidades hacia arriba y hacia la izquierda del pixel. La forma en que las características calculan esta representación intermedia de la imagen es tomando un punto x, y que contiene la suma de los pixeles arriba e izquierda de x, y (ver eq. 4).

$$ii(x, y) = \sum_{\substack{x' \le x \\ y' \le y}} i(x', y') \quad (4)$$

Donde $ii(x, y)$ es la imagen integrada y $i(x, y)$ es la imagen original. Cada parte se calcula con las siguientes recurrencias (eq 5):

$$\begin{aligned} s(x, y) &= s(x, y-1) + i(x, y) \\ ii(x, y) &= ii(x-1, y) + s(x, y) \end{aligned} \quad (5)$$

Donde $s(x, y)$ es el renglón con la suma acumulada, $s(x-1, y) = 0$ y $ii(-1, y) = 0$) la imagen integrada puede ser calculada en un solo paso sobre la imagen original. Se cabe mencionar que si modificamos la implementación actual del funcionamiento de la librería podría ser tener mayor velocidad por recorrer menos veces la imagen de entrada (Viola & Jones, 2004).

Esto permite que el acceso a el promedio de las regiones se obtenga a partir de apuntar únicamente a las cuatro esquinas de la región en lugar de tener que recorrer toda la sección (Ver figura 23). La forma en que se calcula es parecida al cálculo del determinante de una matriz de 2x2 donde sólo se resta la suma de la diagonal menos la suma de la segunda diagonal.

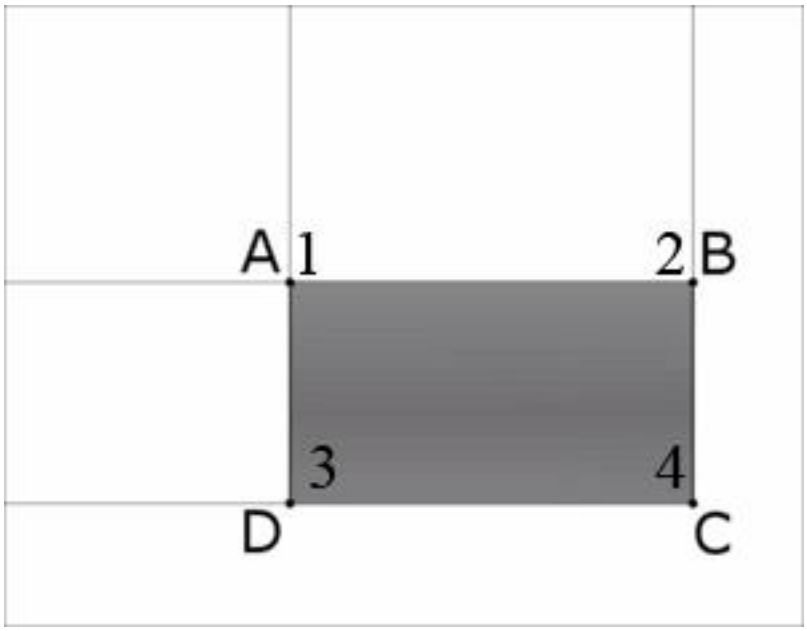

FIGURA 23 EJEMPLO DE REGIÓN

La suma de los pixeles en el rectángulo pueden ser calculados con sólo cuatro referencias a los arreglos. El valor de la imagen en el punto A es 1, en el punto B es 2 lo cual viene de A+B, en el punto D el valor es 3 y se calculó con A+D y en el punto C es 4 calculado de A+B+D+C (ver

ecuación 6). La suma dentro del rectángulo puede ser obtenida de la ecuación es decir, (4+1)-(2+3) (Viola & Jones, 2004).

$$SUM = (A + C) - (B + D) \quad (6)$$

Para la detección específica de rostros se utiliza generalmente en las implementaciones existentes un archivo xml que contiene las características que se encuentran presentes en el rostro, por ejemplo específicamente donde los pómulos son más brillosos que las cuencas oculares y la nariz también con más brillo que las cuencas de los ojos y la frente siendo una de las porciones más brillantes de la cara. Estas comparaciones son hechas miles de veces en tiempo real y a diferentes escalas, esto es permitido debido a la imagen integral.

4.2.3 Adaboost.

La idea principal en el método de Adaboost se basa en tomar clasificadores débiles para tomar una decisión en conjunto sobre la clasificación. Se le llama clasificador débil a aquel que calcula la respuesta correcta apenas por encima del 51%. Entonces el método de Adaboost consistiría en tener por ejemplo tres clasificadores que sus zonas de clasificación de error no choquen para así incrementar la certeza de clasificación.

Partiendo de esta idea, de manera práctica se restringe el clasificador a un conjunto de funciones que dependan de una sola característica para hacer la discriminación. El clasificador débil se diseña para seleccionar la característica rectangular que mejor separa los candidatos negativos de los positivos. Para cada característica el clasificador determina el umbral óptimo de la función de clasificación para lograr la menor cantidad de ejemplos mal clasificados.

Un clasificador débil $h(x, f, p, \theta)$ con un descriptor (f) un umbral θ y una polaridad p que indica la dirección (ver eq. 7).

$$h(x, f, p, \theta) = \begin{cases} 1, si\ \ pf(x) < p\theta \\ 0,\ \ todo\ lo\ demas \end{cases} \quad (7)$$

En la práctica un clasificador con un solo descriptor no podría decidir teniendo un porcentaje bajo de error. Primero se seleccionan los descriptores de bajo porcentaje de error y en las siguientes iteraciones los descriptores que se seleccionan, incrementan su porcentaje. Este clasificador de un solo descriptor, podría verse cómo si fuera un árbol de decisión de un solo nodo.

4.2.4 Análisis en cascada.

El algoritmo barre la imagen en todas las regiones buscando las características entrenadas que se encuentran en el archivo xml. Deslizándose cierta cantidad de pixeles cada vez por la imagen para volver a hacer la evaluación de las características tratando de eliminar todas las potenciales caras falsas. Esto lo repite en diferentes escalas, que se encuentren en los rangos deseados. Eliminando cerca del 50% de los candidatos falsos pero manteniendo el 99% de los rostros.

Los clasificadores más simples son los que se utilizan para desechar la mayoría de las sub-regiones a verificar antes que los clasificadores más complejos sean llamados para obtener positivos falsos.

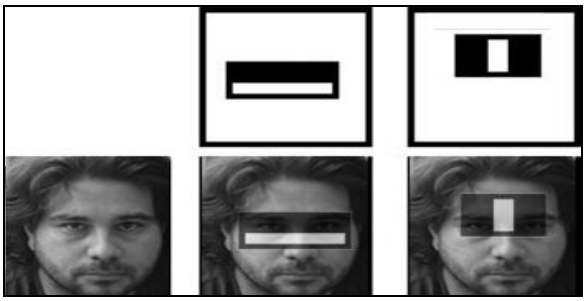

FIGURA 24 CLASIFICADORES DE DOS DESCRIPTORES

Comenzando con el clasificador fuerte de dos descriptores (ver figura 24), un filtro efectivo de rostros se puede obtener de ajustar el umbral del clasificador para minimizar los falsos negativos. El umbral inicial de AdaBoost $\frac{1}{2}\sum_{t=1}^{T}\alpha_t$ está diseñado para obtener la menor tasa de error con los datos de entrenamiento.

El rendimiento de detección de la clasificación de los dos descriptores es bastante aceptable como sistema de detección de rostros. Aun así el clasificador puede reducir significativamente el número de sub-regiones que necesita procesar con pocas operaciones:

1. Evaluar los descriptores de la región (requiere de 6 a 9 referencias por descriptor).
2. Calcular los clasificadores débiles para cada descriptor (requiere una operación con un umbral por descriptor)
3. Combinar los clasificadores débiles (requieren una multiplicación por descriptor, una suma y finalmente un umbral).

Con esto, el sistema puede rechazar casi de inmediato una región y ahorrar cálculos.

4.2.5 Implementación de la detección de rostro.

En este proyecto, los rangos de las características para la detección se redujeron al mínimo para ser más veloces determinando únicamente tamaños que se encontrarían entre el rango de un metro y un metro y medio de la cámara a utilizar. De esta manera la búsqueda en la imagen se reduce lo suficiente para automáticamente buscar los rostros de un tamaño específico (ver Figura 25).

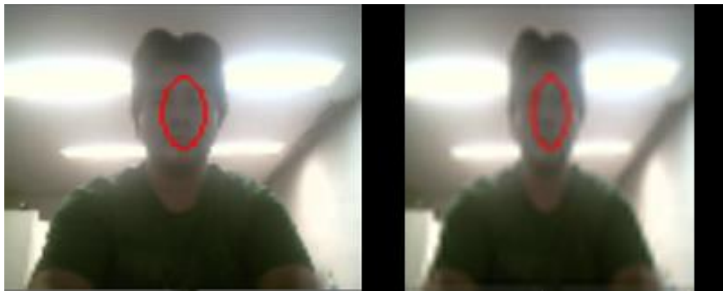

FIGURA 25 IMAGEN DE TAMAÑO ORIGINAL (IZQUIERDA) E IMAGEN REDUCIDA (DERECHA)

La imagen recibida de la cámara se reduce directamente a un tamaño de 130 x 95. Las características de tipo Haar rotadas no son tomadas en cuenta (ver Figura 14). Todo esto se hizo para agilizar la mayor velocidad de procesado posible. También el detector se detiene al momento de encontrar una sola cara. Para detenerla en la búsqueda de más candidatos en la imagen.

Es muy lento reconocer el rostro, aun con los ajustes, el retraso se incrementa hasta dos segundos rápidamente y se continúa acumulando. Para dar solución al problema se generó un intercalado entre las imágenes a reconocer (cada imagen impar pasa por el proceso de reconocimiento y cada imagen par devuelve el mismo valor de posición y reconocimiento que la imagen anterior), esto recuperó la velocidad y permitió incrementar el tamaño de la imagen en la

que se reconocen los rostros, lo cual también fue benéfico para aumentar la calidad de reconocimiento y las distancias soportadas.

4.3 Detección del movimiento

Se utilizó la técnica de diferencia de cuadros para comenzar a filtrar. Esto genera una imagen en la cual es "visible" únicamente el movimiento y las partes en diferentes posiciones con respecto al instante anterior.

Para eliminar la mayoría del ruido *natural* de entrada a la cámara como las vibraciones del color o de la iluminación. Cada imagen de entrada es filtrada mediante una matriz de convolución de 3x3 para promediar cada pixel con sus vecinos. Si recorremos la imagen a lo alto *H* y a lo ancho *W* en cada columna *x* y renglón *y*, los pixeles *P(x,y)* se multiplican por $\frac{1}{9}$ y se suman para dar valor al nuevo pixel *P'(x,y) (ver eq. 8)*.

$$\boldsymbol{P'(x,y) = \sum_{\substack{0 \le x \le W \\ 0<y<H}} P(x,y) * \frac{1}{9}} \quad \textbf{(8)}$$

Después de filtrar el ruido inicial, el método comienza por verificar si se encuentra un rostro en la escena, si es así, comienza a restar los cuadros actuales $f(t)$ y $f(t_{-1})$ para obtener el cuadro de diferencias $D(t)$ de la siguiente manera (*ver eq.* 9):

$$D(t) = f(t) - f(t_{-1}) \quad (9)$$

Con esto, obtenemos la diferencia de la imagen y podemos aislar los cambios en el tiempo o acumular las diferencias. (Ver Figura 26).

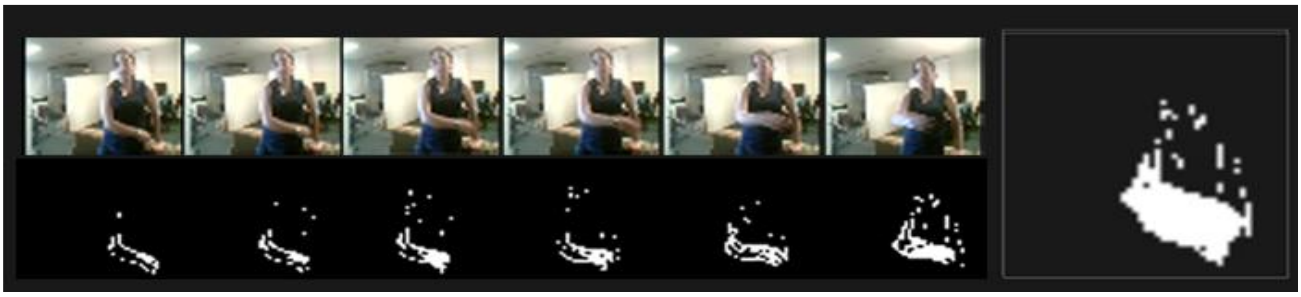

FIGURA 26 DETECCIÓN DEL MOVIMIENTO Y DIFERENCIAS ACUMULADAS

4.4 Reconocimiento del movimiento

Para detectar el movimiento principal deseado, se calculan los histogramas horizontales y verticales a partir de la matriz binarizada de la diferencia de los cuadros restados. La binarización de la imagen esta explicada en el punto 3.3 y la generación de los histogramas horizontales y verticales de cada momento t al iniciar las diferencias de imagen.

Estos dos histogramas generan un pico en la parte con mayor movimiento detectado, esto se debe a la naturaleza de la cercanía con la cámara. El movimiento más próximo a la lente queda registrado como la diferencia más grande (HOO et al., 2005).

En los histogramas sucede el segundo filtro del movimiento. Cada uno al ser calculado junto con la imagen es promediado para suavizar los posibles saltos del movimiento. Tenemos cada histograma *H* como una lista de tamaño 50 y cada valor *i* en la lista del histograma se promedia con sus vecinos de la siguiente manera *(ver eq. 10)*.

$$H'^{(i)} = [H(i-1) + H(i) + H(i+1)] * \frac{1}{3} \quad \textbf{(10)}$$

Para encontrar el movimiento, cruzamos los puntos más altos de ambos histogramas $H_x$de las columnas y H_y de los renglones y el punto $P(H_x(Max), H_y(Max))$ resultante será una parte muy próxima a la mano que se encuentra en movimiento para indicar un gesto (ver Figura 27).

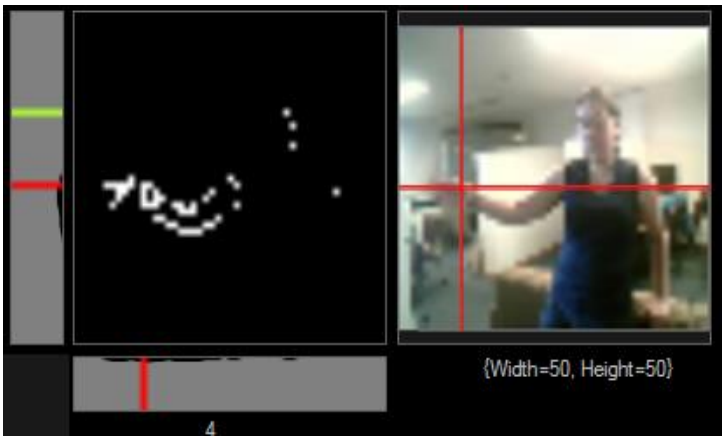

FIGURA 27 INTERSECCIÓN DE PUNTOS MÁXIMOS EN HISTOGRAMAS

4.5 Definición de umbrales

Al principio no era posible tener un punto de comparación para reconocer el gesto. Efectivamente se podía clasificar el movimiento pero generaba la misma cantidad de falsos positivos que correctos positivos, esto debido a la falta de una regla que pudiera tomar una decisión para clasificar el movimiento como gesto. En la clasificación basada en definición del esqueleto, la clasificación del gesto se determina a partir de la comparación del movimiento y las posiciones de la muñeca y el codo o los hombros del cuerpo detectado (Celebi, Aydin, Temiz, & Arici, 2013).

Tomando la misma idea, se utiliza el rostro no sólo para determinar si hay un usuario para analizar su movimiento, también se usa para tener una referencia a la cual comparar el movimiento y que se encuentra en relación con el usuario sin importar su posición exacta.

A partir de esto se determinan tres umbrales para crear tres diferentes reglas que determinen si el movimiento detectado es un gesto que indique, izquierda, derecha o arriba. Al tener la posición de la cara de la persona, se elimina todo movimiento por encima de esta para limpiar la imagen binaria y filtrar ruido. Esto permite que el usuario pueda moverse más natural en su misma posición sin que se detecte el movimiento o se clasifique cierta cantidad detectada erróneamente como gesto. Parte del movimiento involuntario que la mayoría de los gestos reciben como ruido es filtrado de esta manera.

Se fijan tres umbrales finales a partir de la cara, uno es la parte media del rostro en torno con su altura. Este umbral sirve para detectar si el movimiento analizado se debe clasificar como el gesto de *subir*. Si el rostro se detecta en el punto $P(x, y)$ con una altura *H* y un ancho *W* y *P(x,y)* es el rectángulo donde comienza la cara en la esquina superior izquierda en la imagen. Entonces, el punto medio del rostro será dado por la ecuación 11.

$$\boldsymbol{U_1 = P(x + \frac{W}{2}, y + \frac{H}{2})} \quad \textbf{(11)}$$

El segundo umbral se encuentra recorrido a la derecha del rostro a una distancia del tamaño de la cara multiplicado por 2.5, esta relación es una aproximación a la posición de la mano apuntando hacia la derecha del cuerpo rebasando su propio codo y hombro (ver eq. 12).

$$\boldsymbol{U_2 = P(x - W * 2.5, y + \frac{H}{2})} \quad \textbf{(12)}$$

El último umbral se encuentra dónde termina el rostro del lado izquierdo, dado por el mismo punto $P(x, y)$ y la suma de su ancho W. (ver eq. 13)

$$\boldsymbol{U_3 = P(x + W, y)} \quad \textbf{(13)}$$

En la Figura 28 podemos ver una imagen de los umbrales marcados como se mencionan anteriormente.

FIGURA 28 UMBRALES

Si el movimiento dado por el punto $P(H_x(Max), H_y(Max))$ cruza alguno de estos umbrales en sus dos puntos a la vez dado por el momento $M(t)$ y $M(t-1)$ se activa el gesto.

4.7 Clasificando el movimiento

A partir del cruce generado por los histogramas, se guarda la posición del movimiento $M_{t-1}\ (P\left(H_x(Max), H_y(Max)\right))$ del cuadro anterior y del cuadro actual $M_t\ (P\left(H_x(Max), H_y(Max)\right))$, esto nos genera un vector con dirección del movimiento. En la figura 29 podemos ver en una cruz amarilla dónde está detectado en una imagen de 50x50 pixeles el centro de un rostro, con unas líneas rojas podemos nuevamente ver dónde se ubicarían los 3

umbrales con respecto al rostro detectado, después en color blanco podemos observar el movimiento registrado en la imagen de diferencias y binarizado y en círculos de color rojo y verde el vector de movimiento, donde el círculo rojo representa la posición anterior y el círculo verde la posición actual y su dirección.

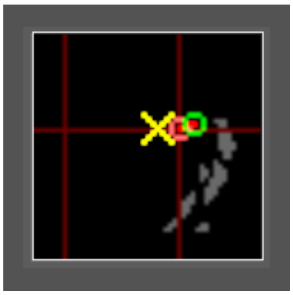

FIGURA 29 VECTOR DE MOVIMIENTO

Cuando el vector v completo atraviesa uno de los tres umbrales se registra como gesto detectado, dependiendo cuál de los tres umbrales U_1, U_2 y U_3 descritos en 4.5 nos genera la clasificación del gesto dado el movimiento del vector $G(v)$ *(ver eq. 14).*

$$\mathbf{G}(v)\begin{cases} \boldsymbol{UP}, & \boldsymbol{v > U_1} \\ \boldsymbol{LEFT}, & \boldsymbol{v < U_2} \\ \boldsymbol{RIGHT}, & \boldsymbol{v > U_3} \end{cases} \tag{14}$$

Teniendo como prioridad el gesto de arriba del rostro, esto debido a que se compara del histograma vertical la posición donde comienza el movimiento de manera vertical, en segundo lugar hacia afuera y en tercer lugar hacia adentro o en este caso, izquierda y derecha respectivamente. En la figura 30 podemos ver el movimiento encontrado con los histogramas, el rostro y el vector generado. Las líneas rojas indican los límites marcados de acuerdo con la ubicación del rostro.

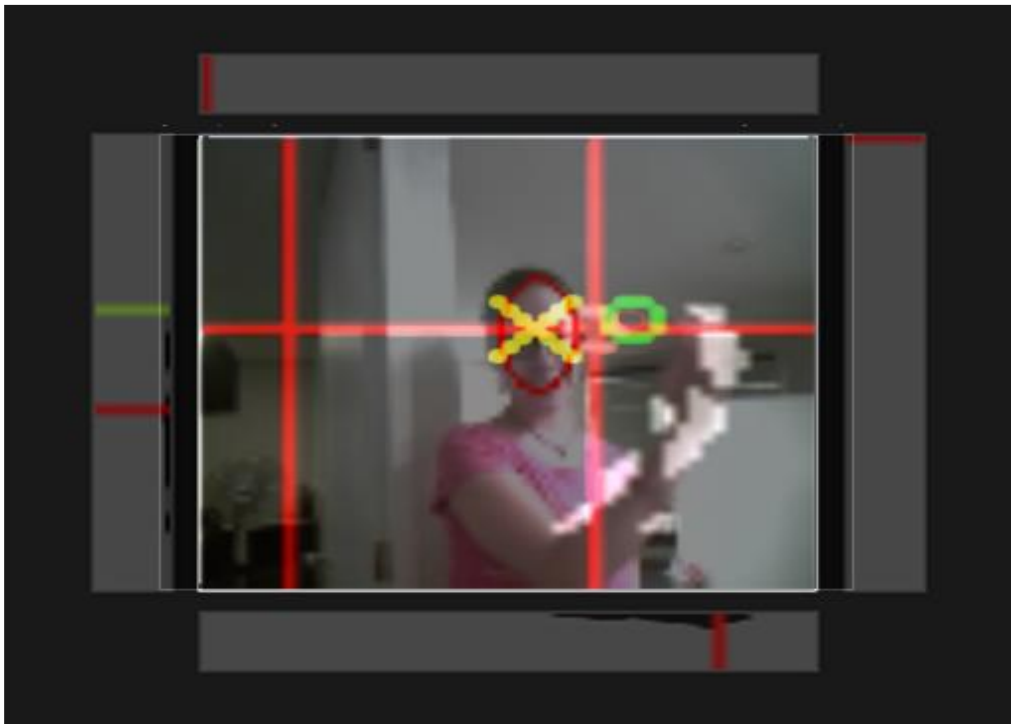

FIGURA 30 RESULTADO FINAL

4.7 Condiciones del método

Al estar condicionado a la detección del rostro, la técnica necesita tener controlada la iluminación de manera frontal al usuario. La cantidad de luz no debe quemar el rostro del usuario pero tampoco deberá dominar la iluminación trasera, esto podría exagerar las sombras en el rostro y como consecuencia no funcionar el módulo de reconocimiento del rostro.

Por otro lado, los movimientos que la gente genere necesitan ser de una velocidad preferentemente alta para exagerar el cambio en los cuadros entre una imagen y la siguiente. Si el cambio es suficientemente pequeño no se notará y el sistema no podrá reconocer algún tipo de movimiento.

4.7.1 Posición de la cámara

Dada la posición de la cámara, el ángulo afecta a la detección del movimiento. Dado que la premisa es que el objeto más cercano será el que genere un movimiento más grande debido a la diferencia en la distancia que genera (HOO et al., 2005). Si la cámara se encuentra inclinada en aproximadamente 30 grados hacia arriba detectando el rostro, el movimiento generado por la mano comienza a ser constantemente reconocido como gesto de *arriba.*

4.7.2 Rangos de movimiento

Con la implementación y el comportamiento del sistema podemos encontrar diferentes casos y variantes a cada fase del reconocimiento de los gestos. Con estos valores iniciales vemos para donde podríamos mover los valores para comenzar a expandir la teoría e implementar más gestos y mejorar el desempeño del mismo. A continuación se presentará la evaluación del método seleccionado.

Acabamos de describir el método que se implementó de manera eficaz para registrar movimiento y poderlo clasificar como un gesto dinámico mímico. A continuación se presentarán los resultados obtenidos de un proceso de pruebas con diferentes sujetos de diferentes edades, estaturas y contextos.