Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

Tipo de artículo: Artículo original
Temática: Reconocimiento de patrones
Recibido: 01/10/2015 | Aceptado: 20/12/2015

Detectores espacio-temporales para la detección de rostros en video

Spatio-temporal detectors for face detection in video

Yoanna Martínez-Díaz[1*], Noslen Hernández[2], Heydi Méndez-Vázquez[1]

[1]Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV). Avenida 7ma A ♯ 21406 % 214 y 216, Siboney, Playa, P.C. 12200, Habana, Cuba. {ymartinez,hmendez}@cenatav.co.cu
[2]Pontifícia Universidade Católica do Rio de Janeiro, Brasil. nhernandez@gmail.com

[*]Autor para correspondencia: ymartinez@cenatav.co.cu

Resumen
La detección de rostros es el primer paso en muchas aplicaciones de video como la video vigilancia, el análisis de expresiones faciales, el seguimiento y el reconocimiento de rostros. Varios algoritmos han sido propuestos para llevar a cabo esta tarea; sin embargo, la mayoría de ellos se basan en técnicas para imágenes fijas y no consideran la información espacio-temporal existente en un video. En este trabajo se desarrollan dos detectores de rostros espacio-temporales, los cuales son evaluados en la base de datos *YouTube Faces*. Los resultados alcanzados son comparados con los obtenidos por dos detectores que se basan únicamente en la información espacial.

Palabras claves: detección de rostros, video, representación espacio-temporal

Abstract
Face detection is the first step in many video applications such as video surveillance, facial expression analysis, face tracking and face recognition. Several algorithms have been proposed to this task, but most of them are based on techniques for still images, not considering the spatio-temporal information available in a video. In this paper two spatio-temporal face detectors are developed and evaluated on the challenging YouTube Faces database. The obtained results are compared with those obtained by two frame-based approaches.

Keywords: *face detection, video, spatio-temporal representation*

Introducción

La detección de rostros se ha convertido en una de las áreas más investigadas por la comunidad científica (ZHANG and ZHANG, 2010), considerándose en muchos sistemas automáticos de video como el primer paso a realizar. Sin embargo, la mayoría de los procesamientos posteriores suponen que ya el rostro ha sido detectado.

Diversos métodos han sido propuestos en la literatura para resolver el problema de la detección de rostros en video. Entre los enfoques existentes, aquellos que resuelven el problema cuadro a cuadro han sido uno de los más usados (FROBA and KUBLBECK, 2004; WAEL, 2011). Estos métodos han sido desarrollados originalmente para detectar los rostros en imágenes fijas, por lo que no tienen en cuenta la correspondencia temporal existente entre los cuadros consecutivos de un video. En este caso, cada cuadro del video es analizado de manera independiente, como si se tratara de una nueva imagen, lo que cual hace más lento el proceso. Por otra parte, se han utilizado algoritmos de detección de movimiento con el fin de detectar el rostro (NASCIMENTO and MARQUES, 2006). Sin embargo, la mayoría de estos métodos solo dan buenos resultados en entornos estáticos o donde el fondo cambia lentamente.

A pesar de los resultados alcanzados mediante el uso de descriptores espacio-temporales en diferentes aplicaciones de análisis facial (BARR et al., 2012), se le ha prestado muy poca atención a su uso en el contexto de la detección de rostros en video. Representaciones como los volúmenes de patrones binarios locales (VLBP, por sus siglas en inglés) (ZHAO and MATTI, 2007), el conjunto extendido de los VLBP (EVLBP) (HADID and PIETIKÄINEN, 2009) y recientemente, los volúmenes de características ordinales estructuradas (VSOF, por sus siglas en inglés) (MENDEZ-VAZQUEZ et al., 2013) han sido usadas satisfactoriamente en aplicaciones como el reconocimiento de rostros, de expresiones faciales y la clasificación de género; mostrando además, los beneficios de integrar la coherencia temporal y espacial de la apariencia del rostro.

Recientemente en MARTINEZ-DIAZ et al. (2013) fue propuesto un nuevo enfoque para la clasificación en rostro/no-rostro de secuencias de video. En este trabajo se muestra que el uso del descriptor espacio-temporal EVLBP mejora la eficacia del clasificador *Adaboost* y se obtienen mejores resultados en comparación con tres enfoques que solo utilizan la información espacial. Sin embargo, los autores de este trabajo solo se centran en decidir si una secuencia dada es un rostro o no, mientras que en un escenario real todo el video debería poder ser analizado y cada zona candidata clasificada como rostro/no-rostro; dando como salida final la posición de cada uno de los rostros detectados en el video.

En el presente trabajo se proponen dos detectores espacio-temporales para la detección de rostros en secuencias de video. Primero, teniendo en cuenta los resultados preliminares obtenidos en MARTINEZ-DIAZ et al. (2013), se desarrolló un detector basado en el descriptor EVLBP. Segundo, motivados por las mejoras alcanzadas por el descriptor VSOF sobre el descriptor EVLBP en el reconocimiento facial en video, decidimos extender su aplicación al caso de la detección de rostros; creando así un detector de rostros basado en dicho descriptor. Por último, el desempeño de ambos detectores es evaluado y comparado con otros detectores diseñados para imágenes fijas, con el objetivo de mostrar las ventajas de nuestra propuesta.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

Detectores de rostros espacio-temporales

Los detectores de rostros que se proponen en este trabajo utilizan: (1) un descriptor espacio-temporal para codificar tanto la información espacial como la temporal de cuadros consecutivos en un video; (2) un algoritmo *boosting* para seleccionar y aprender de manera automática los rasgos más discriminativos y (3) un esquema de cascada de clasificadores para acelerar el proceso de la detección.

La mayoría de los descriptores espacio-temporales propuestos en la literatura son extensiones al dominio del video, de descriptores basados en la apariencia local. Por ejemplo, el VLBP (ZHAO and MATTI, 2007) es la primera extensión del descriptor LBP, el cual trata una secuencia de video como un prisma rectangular, comparando cada píxel no solo con sus píxeles vecinos en el dominio espacial sino también con los de sus cuadros más cercanos. Luego, en HADID and PIETIKÄINEN (2009), se propone el conjunto extendido del VLBP (EVLBP), siendo este un descriptor más flexible ya que permite usar varios parámetros de configuración como diferentes radios, número de puntos de muestreo e intervalos de tiempo. El operador EVLBP para cada píxel en cada cuadro se obtiene de la siguiente manera:

$$EVLBP_{L,(P,Q,S),R} = \sum_{m=0}^{M-1} s(I_{t,m} - I_{t_c,c})2^m, \tag{1}$$

donde t_c corresponde al cuadro del píxel del centro c y t es cada cuadro usado en el proceso de codificación; L es el intervalo de tiempo entre los cuadros codificados, de modo que $t = t_c - 2L, t_c - L, t_c, t_c + L, t_c + 2L$; R es el radio para la selección de los píxeles vecinos; $M = P + 2Q + 2S$ es número total de píxeles codificados, elegidos de la siguiente forma: P píxeles del cuadro t, Q del cuadro $\pm t$ y S del cuadro $\pm 2t$; $I_{t_c,c}$ es el valor de intensidad del píxel del centro y $I_{t,m}$ del píxel m en el cuadro t; $s\{f\} \in \{0, 1\}$ es un indicador booleano de la condición f.

Recientemente en MENDEZ-VAZQUEZ et al. (2013), los autores inspirados en el descriptor EVLBP, propusieron el descriptor VSOF, el cual mantiene la misma configuración flexible lo que en lugar de comparar directamente los valores de los píxeles, compara los valores promedios de regiones. El tamaño o escala de estas regiones puede ser diferente para cada configuración. De esta forma, estructuras más complejas pueden ser representadas. El descriptor VSOF, a una escala N dada (tamaño de las regiones $N \times N$), se puede obtener reescribiendo la eq.(1) de la siguiente manera:

$$VSOF_{L,(P,Q,S),R,N} = \sum_{m=0}^{M-1} s(g_{t,m} - g_{t_c,c})2^m, \tag{2}$$

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

donde c es la posición central de la región del centro y g_i es la intensidad promedio de la región i. En la Figura 1 se muestra un ejemplo de un código VSOF obtenido usando $L = 2, P = 4, Q = 3, S = 1, R = 3, N = 3$.

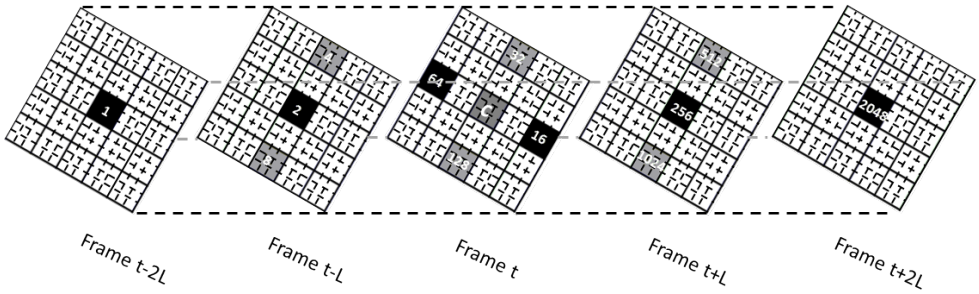


Figura 1. Ejemplo del proceso de codificación del operador VSOF. En este caso $VSOF_{2,(4,3,1),3,3} = 110001011001 = 3161$

Note que, el descriptor EVLBP es un caso del VSOF cuando $N = 1$. Para grandes escalas, o sea, $N > 1$, la imagen integral es usada para calcular los valores de intensidad promedio de cada una de las regiones, lo cual reduce el costo computacional. Tanto para el descriptor EVLBP como para el VSOF, mediante el uso de diferentes configuraciones de parámetros, se obtienen un conjunto extenso de códigos (conjunto exhaustivo). En trabajos previos (HADID and PIETIKÄINEN, 2009; MARTINEZ-DIAZ et al., 2013; MENDEZ-VAZQUEZ et al., 2013) se han utilizado la distribución de estos códigos (histogramas) para describir las secuencias de video; donde cada elemento o característica del vector que se obtiene corresponde a un bin del histograma para una configuración dada. En este trabajo vamos a explorar además, el uso directo de los códigos como representación, tanto para el descriptor EVLBP como para el VSOF. En este caso, cada elemento del vector que se obtiene representa un código con una configuración en una posición dada.

Un aspecto importante en este tipo de descriptores para el caso de la detección, es el número de de cuadros (M) a usar en la codificación; ya que si usáramos muchos cuadros podríamos estar incluyendo en la codificación información adicional como es el caso del fondo. Sin embargo, si se utilizara un número muy pequeño de cuadros se podría perder información espacio-temporal discriminativa, y se requeriría más tiempo para procesar el video completo. Por tanto, es necesario encontrar un compromiso entre asegurar que solo la información del rostro sea representada, minimizando la variabilidad que se codifica, y maximizar el número de cuadros a utilizar.

Como se mencionó anteriormente, tanto el descriptor EVLBP como el VSOF permiten utilizar varios parámetros de configuración, lo que genera un gran conjunto de rasgos. Con el objetivo de seleccionar aquellos rasgos más discriminativos y así eliminar información redundante, aplicamos un algoritmo *boosting*. En la literatura se han propuesto diferentes variantes de algoritmos *boosting* para la detección de rostros (ZHANG and ZHANG,

2010). En este trabajo se utiliza el algoritmo *GentleBoost* (FRIEDMAN et al., 1998) ya que es fácil de implementar y ha mostrado mejores resultados que otras variantes. Para ambos descriptores (EVLBP y VSOF), en el caso de la representación por histogramas, empleamos como clasificadores débiles los *regression stumps*; mientras que para el caso de la representación basada códigos, utilizamos árboles de regresión de múltiples ramas; ya que estos son características no métricas y no es posible utilizar funciones de umbralización. Para la construcción de la cascada de clasificadores *boosting* se siguió el esquema de Viola y Jones (VIOLA and JONES, 2004).

Para estos detectores espacio-temporales, en lugar de usar una ventana deslizante como para el caso de las imágenes, vamos a utilizar un prisma o cubo deslizante de profundidad M; para buscar en todo el video a distintas escalas y ubicaciones, determinando cuando una región se corresponde o no con un rostro. De esta forma, no necesitamos escanear cada cuadro del video sino solo cada M cuadros consecutivos, lo que reduce el tiempo de cálculo comparado con los algoritmos que realizan la detección cuadro a cuadro. Así, cuando un cubo es clasificado como rostro, vamos a tener la misma detección en los M cuadros que pertenecen al cubo detectado. De igual manera cuando el detector falle u obtenga un falso positivo vamos a perder el rostro o a mantener el falso positivo en los M cuadros correspondientes.

Por otra parte, se propuso un método de pos-procesamiento para unir múltiples detecciones que se encuentren sobre un mismo objeto. Para esto, agrupamos todas las detecciones obtenidas en un cubo basados en una medida de disimilitud, la cual tiene en cuenta dos aspectos fundamentales: (1) el área compartida por dos detecciones, no solo en términos de cantidad sino también en términos de significancia de la región de intersección y (2) la diferencia de escala entre las dos detecciones. Para garantizar el primer aspecto tenemos una distribución de importancia de la región (G_s) para la plantilla de detección en cada escala s; tal que mientras más cercana esté la región al centro de la plantilla, mayor significancia tendrá esta (para esto se usó una densidad gaussiana bivariada). De esta forma se tiene en cuenta la ubicación del área de intersección dentro de cada detección. La medida de disimilitud usada, puede ser formalizada para dos detecciones d_i y d_j como:

$$D(d_i, d_j) = 1 - \left(1/2(G_{s_{d_i}}(I(d_i, d_j)) + G_{s_{d_j}}(I(d_i, d_j)))p(s_{d_i}, s_{d_j})\right),$$

donde s_{d_i} y s_{d_j} son las escales de d_i y d_j, respectivamente. $G_{s_{d_i}}$ y $G_{s_{d_j}}$ dan la contribución del área de intersección $I(d_i, d_j)$; mientras que $p(s_{d_i}, s_{d_j})$ penaliza la diferencia de escala, tal que mientras mayor sea esta diferencia mayor será la penalización y por consiguiente, menor será el valor de $p(s_{d_i}, s_{d_j})$. El intervalo de valores de p oscila entre $(0-1]$. Note que, la disimilitud $D(d_i, d_j)$ toma valor 0 cuando d_i y d_j concuerdan completamente, es decir, son iguales y toma valor 1 cuando d_i y d_j no se intersecan. El algoritmo de agrupamiento utilizado fue el *average-linkage*. Una vez que se obtienen los grupos, aquellos con menos de tres detecciones

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

fueron eliminados; mientras que de cada uno de los grupos restantes solo se consideró la detección que tuviera el mayor valor de confianza, el cual está dado por el valor del clasificador.

Resultados y discusión

En el caso de la detección de rostros en videos sigue siendo difícil comparar diferentes algoritmos, debido a la falta de un banco de prueba que permita usar protocolos de evaluación comunes. Además, la mayoría de las bases de datos de videos existentes no fueron diseñadas específicamente para esta tarea y no reflejan algunos aspectos que se manifiestan en los escenarios reales. Por estas razones en este trabajo decidimos realizar nuestros experimentos en la base de datos *YouTube Faces* (WOLF et al., 2011); la cual contiene 3425 videos de 1595 personas con distintas expresiones, condiciones de iluminación, poses, resoluciones y fondos. En esta sección se realizaron dos experimentos fundamentales. En el primer experimento se comparan los descriptores EVLBP y VSOF y en el segundo experimento se describe la construcción de los detectores propuestos y se evalúa su desempeño.

Comparación de las representaciones espacio-temporales

En este experimento se compara el poder discriminativo de la representación por códigos y por histogramas para ambos descriptores (EVLBP, VSOF). Por lo que solo nos vamos a centrar en la clasificación en rostro/no-rostro, o sea, en decir si una muestra dada es o no un rostro.

Con este propósito, seleccionamos prismas rectangulares (cubos) de muestras positivas (rostros) y muestras negativas (no-rostros) usando $M = 14$. El valor de M fue seleccionando teniendo en cuenta el trabajo realizado en MARTINEZ-DIAZ et al. (2013). Para los cubos positivos se utilizó la anotación del rostro en cada cuadro del video, proporcionada por la base de datos. Dado que la posición del rostro en cada cuadro no es la misma, el volumen real que se forma con las anotaciones de los 14 cuadros consecutivos no forma necesariamente un prisma rectangular. Por tanto, para poder capturar tanto como fuese posible los desplazamientos reales del rostro, seleccionamos el prisma rectangular de mayor intersección con el volumen real. En el caso de las muestras negativas, los cubos se extrajeron aleatoriamente del fondo de los videos donde no hubieran rostros.

En total se seleccionaron 40000 cubos positivos (rostros) y 70000 cubos negativos (no-rostros) de tamaño $40 \times 40 \times 14$. De estos, se tomaron 10000 rostros y 10000 no-rostros para el entrenamiento y el resto para la prueba. Se usó como clasificador el *GentleBoost* con 100 clasificadores débiles. Los descriptores EVLBP y VSOF fueron extraídos como se explicó en la sección anterior, usando diferentes configuraciones de parámetros.

En la Tabla 1 se muestran algunas características de las representaciones utilizadas tales como el tamaño de

los vectores de rasgos y el tiempo promedio de extracción (en segundos) de estos. Además se muestran los resultados de la clasificación en términos de tasa de de Falsos Negativos (FN) y Falsos Positivos (FP). Como se puede observar el descriptor VSOF obtiene mejores resultados que el EVLBP tanto usando la presentación por códigos como por histogramas. Para ambos descriptores el uso de los códigos directamente supera la representación por histogramas tanto en eficacia como en eficiencia; a pesar de que el tamaño de los vectores de estos últimos es menor.

Teniendo en cuenta los resultados obtenidos, construimos nuestros detectores basados en la representación por códigos.

Tabla 1. Comparación de los descriptores espacio-temporales.

	Tamaño del vector	**Tiempo de extracción**	**FN (%)**	**FP (%)**
Códigos-EVLBP	78720	0.012	1.45	1.60
Hist-EVLBP	2048	0.055	3.10	6.14
Códigos-VSOF	157056	0.031	1.15	1.46
Hist-VSOF	6144	0.121	1.92	4.99

Evaluación de los detectores

En este experimento los dos detectores de rostros propuestos son diseñados y creados usando el esquema de Viola y Jones. El procedimiento para ambos detectores es el mismo solo se van a diferenciar en los descriptores usados para la representación.

Para el entrenamiento de la cascada se utilizaron 21350 cubos de rostros del experimento anterior y 21350 cubos de no-rostros, los cuales fueron seleccionados de 300 videos descargados de internet que no contienen ningún rostro. Después que cada etapa es entrenada, se utiliza la estrategia *bootstrap* para obtener muestras negativas mal clasificadas por la cascada entrenada hasta ese momento; las cuales serán usadas en el entrenamiento de la próxima etapa. La tasa máxima de falsas alarmas y la tasa mínima de detección establecidas fueron de 0.5 y 0.995, respectivamente.

Como resultado obtuvimos dos detectores de 14 etapas cada uno. El detector-EVLBP con un total de 114 rasgos y el detector-VSOF con 98 rasgos. Ambos fueron evaluados y comparados con dos detectores cuadro a cuadro, también de 14 etapas: el detector-Haar con 180 rasgos en total y el detector-LBP con 82 rasgos. Estos dos últimos fueron creados usando la implementación de la *OpenCV*.

Para la evaluación se utilizaron 133 videos de la base de datos *YouTube Faces*, diferentes a los empleados para el entrenamiento. Esta base de datos solo proporciona la anotación de un rostro; sin embargo existen

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

algunos videos que presentan más de un rostro. Por lo que nosotros, manualmente, anotamos estos videos con el objetivo de poder evaluar nuestros detectores ante la presencia de múltiples rostros.

Los resultados obtenidos se muestran en la Tabla 2 en términos de tasa de detección y falsas alarmas por cuadro. Una respuesta de un detector se considera una detección correcta basado en el siguiente criterio:

$$score = \frac{area(d_i \cap g_i)}{area(d_i \cup g_i)}, \tag{3}$$

donde d_i es la detección obtenida y g_i corresponde a la anotación verdadera. Finalmente, si el $score > 0{,}5$, la detección obtenida se considera como correcta.

De la tabla se puede observar que el detector-VSOF supera al resto de los detectores con menor número de falsas alarmas por cuadro. A pesar de que los detectores EVLBP y LBP logran similares tasas de detección, el detector-LBP produce muchas más falsas alarmas. En la Figura 2 se muestran algunos resultados de los detectores en 5 cuadros representativos de dos secuencias de video. Como se puede apreciar, usando la información espacio-temporales se logra un mejor comportamiento con menor cantidad falsas alarmas.

Tabla 2. Resultados obtenidos por los distintos detectores.

	Tasa de detección (%)	**Falsas Alarmas por cuadro**
Detector-Haar	77.78	3.48
Detector-LBP	82.40	2.23
Detector-EVLBP	82.38	0.98
Detector-VSOF	**87.88**	**0.42**

Conclusiones

En este trabajo se propusieron dos detectores de rostros, los cuales se basan en la información espacial y temporal disponible en un video. Para ello se utilizaron descriptores espacio-temporales que permiten codificar y representar los patrones del rostro en un conjunto de cuadros consecutivos. Los experimentos realizados en la base de datos *YouTube Face* mostraron que tanto para el descriptor EVLBP como para el VSOF la representación basada en códigos es más eficiente y más discrimnativa que la representación por histogramas. Además, los detectores desarrollados mostraron mejor comportamiento y resultados más exactos que detectores que solo consideran la información espacial. Nuestro trabajo futuro estará enfocado en el desarrollo de un método que nos permita asociar las detecciones obtenidas para así construir la trayectoria de cada uno de los rostros presentes en un video. Otra posible linea pudiera ser la exploración o creación de nuevas representaciones

espacio-temporales.

Agradecimientos

El segundo autor de este trabajo es financiado por FAPERJ/CAPES (E45/2013).

Referencias

BARR, J. R., BOWYER, K. W., FLYNN, P. J., and BISWAS, S. (2012). Face Recognition from Video: a Review. *IJPRAI*, 26(5).

FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R. (1998). Additive Logistic Regression: a Statistical View of Boosting. *Annals of Statistics*, 28:2000.

FROBA, B. and KUBLBECK, C. (2004). Face tracking by means of continuous detection. In *IEEE Conf. Comput. Vision Pattern Recognition (CVPR) Workshops,*.

HADID, A. and PIETIKÄINEN, M. (2009). Combining appearance and motion for face and gender recognition from videos. *Pattern Recogn.*, 42(11):2818–2827.

MARTINEZ-DIAZ, Y., MENDEZ-VAZQUEZ, H., HERNANDEZ, N., and GARCIA-REYES, E. (2013). Improving faces/non-faces discrimination in video sequences by using a local spatio-temporal representation. In *ICB*, pages 1–5.

MENDEZ-VAZQUEZ, H., MARTINEZ-DIAZ, Y., and CHAI, Z. (2013). Volume structured ordinal features with background similarity measure for video face recognition. In *ICB*, pages 1–6.

NASCIMENTO, J. C. and MARQUES, J. S. (2006). Performance evaluation for object detection algorithms for video surveillance. In *IEEE Transaction on Multimedia*.

VIOLA, P. and JONES, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

WAEL, L. (2011). Co-occurrence of local binary patterns features for frontal face detection in surveillance applications. *EURASIP Journal on Image and Video Processing*, 2011.

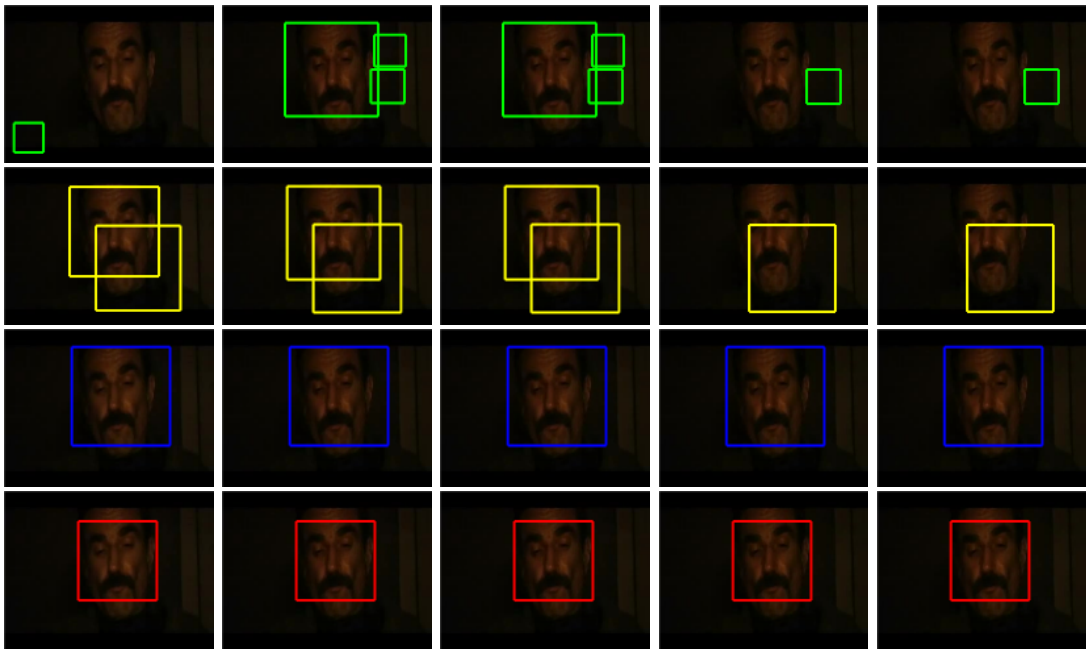
WOLF, L., HASSNER, T., and MAOZ, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*.

ZHANG, C. and ZHANG, Z. (2010). A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research,, Redmond, Washington.

ZHAO, G. and MATTI, P. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6):915 –928.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 205-214

(a) Secuencia 1



(b) Secuencia 2

Figura 2. Resultados de la detección en dos secuencias de video. Detecciones en verde corresponden al detector-Haar, en amarillo al detector-LBP, en azul al detector-EVLBP y en rojo al detector-VSOF.