Algorithm for Early Threat Detection by Suspicious Behavior Representation

D. Martínez, H. Loaiza, and E. Caicedo

***Abstract*—The proposed early detection algorithm is justified because the probability of success to control a criminal activity increases when the response time for generating a warning alarm is reduced. In this paper, a video-based representation model to describe suspicious behavior from elementary actions is proposed. Such behaviors allow detecting potential threats before suspects achieve physical contact with their potential victims. In the algorithm, a novel method to adjust the balance between the anticipation level to threats and the generation of false alerts is introduced. The experimental results obtained from two validation datasets, with attacks to pedestrians and threats against a parked truck, demonstrated the effectiveness of the proposed approach for early threat detection, with performance measures above 90%.**

***Index Terms*—Surveillance, Video Understanding, Suspicious Behavior, Threat Detection.**

I. Introducción

LOS asaltos a peatones son actividades criminales comunes en ambientes urbanos, principalmente en lugares solitarios y/o con baja visibilidad debido a que los asaltantes pueden tomar ventaja para acechar y sorprender a sus víctimas. Dentro de este tipo de escenarios, acciones simples realizadas por algún sujeto tales como cambiar la dirección de su trayectoria, permanecer en espera, agacharse, etc., pueden considerarse normales si se observan de forma aislada. Sin embargo, si un mismo sujeto repite o efectúa varias de estas acciones elementales, entonces su comportamiento podría mostrarse sospechoso. Inspirado en esta idea se presenta un modelo algorítmico jerárquico para representar por medio de eventos simples, los cuales están relacionados entre sí, comportamientos sospechosos de mayor complejidad.

Existe una amplia variedad de trabajos relacionados con sistemas automáticos de vigilancia basados en video. Algunas propuestas tales como [1], [2] se enfocan en la detección de peatones. Otros autores como [3], [4], [5] presentan propuestas para el seguimiento de personas en la escena. En [6] emplean un algoritmo de segmentación para detectar objetos en movimiento en un sistema de video vigilancia para un estacionamiento de vehículos. En [7] los autores proponen un modelo DLS (Domain Specific Language) para configurar e integrar en forma automática, diferentes aplicaciones en sistemas de video vigilancia. Sin embargo, en los trabajos mencionados previamente no se realiza ningún tipo de análisis para la detección o interpretación automática de comportamientos.

D. Martínez, Universidad del Valle, Santiago de Cali, Colombia, e-mail: duber.martinez@correounivalle.edu.co.

H. Loaiza, Universidad del Valle, Santiago de Cali, Colombia, e-mail: humberto.loaiza@correounivalle.edu.co.

E. Caicedo, Universidad del Valle, Santiago de Cali, Colombia, e-mail: eduardo.caicedo@correounivalle.edu.co.

Un avance en esta dirección es el objetivo de otro grupo de trabajos que se centran en la detección automática de amenazas en videos [8]. De estos, la mayor parte se enfocan en la detección de acciones violentas como peleas y agresiones [9]–[11]. Típicamente en este tipo de algoritmos se emplea una ventana espacio-temporal estrecha sobre el evento de interés, a partir de la cual se buscan patrones de movimiento considerados anormales [12], [13]. De esta forma, la detección de la amenaza sólo puede realizarse cuando se presenta la agresión física.

Son relativamente escasos los en detectar de forma temprana una potencial amenaza antes de que ocurra la agresión física. Las propuestas que mejor se aproximan a este enfoque son aquellas que buscan identificar comportamientos específicos que resulten sospechosos, como detección de trayectorias no habituales [14]–[18] o una persona que deambula por la escena. Una limitación en este tipo de enfoques es que sólo identifican la ocurrencia del evento, pero no realiza una cuantificación de la acción que permita comparar eventos del mismo tipo. Por ejemplo, determinar para dos sujetos que deambulan por la escena cuál de ellos representa una mayor amenaza.

A diferencia de propuestas como [9]–[11] donde el interés se centra principalmente en la ventana de tiempo cuando ocurren las acciones violentas, el algoritmo ADTA (Algoritmo para Detección Temprana de Amenazas) propuesto en el presente artículo se enfoca en los eventos previos a dicha ventana de tiempo. De esta forma, es posible generar alarmas tempranas que den aviso de una potencial amenaza antes que ocurra el contacto físico. La generación temprana de una alerta se logra cuando se puede detectar una posible amenaza mientras los atacantes aún se encuentran separados cierta distancia de sus víctimas. Entre mayor sea el intervalo de tiempo entre la detección de una amenaza y la ejecución de la acción criminal, más preventiva y útil será la alerta generada; sin embargo, también se incrementará la probabilidad de generar falsas alertas. Una de las principales características del algoritmo ADTA propuesto es que permite establecer un balance entre la capacidad de una detección temprana y la generación de falsas alarmas mediante la introducción de un parámetro de ajuste. Métodos como [14]–[18] que mediante la detección de trayectorias anormales pueden en algunos casos permitir la generación de alertas tempranas no cuentan con esta capacidad de ajuste, lo que condiciona la confiabilidad de la alerta.

En resumen, las principales contribuciones presentadas en el artículo respecto a trabajos previos incluyen:

- Se propone un novedoso modelo algorítmico para representar comportamientos complejos como una combinación de acciones elementales.

- El modelo propuesto permite cuantificar el nivel de amenaza que representa un conjunto de acciones durante una escena registrada en video.

- El algoritmo ADTA propuesto no se limita en la ventana de tiempo durante la cual ocurre la agresión, por el contrario, reúne la información de las acciones previas y de esta forma permitir detectar una posible amenaza antes de que se produzca un contacto físico entre el agresor y la víctima.

- Se presenta un análisis experimental donde se relaciona la capacidad de una detección temprana respecto a la generación de falsas alarmas.

Para la validación del esquema propuesto se utilizaron dos conjuntos de datos de prueba: el primero de libre acceso y denominado PETS2014 [19], que contiene videos con amenazas contra un vehículo estacionado y contra su conductor. El segundo es un conjunto de datos propio denominado UVS-dataset que contiene escenas dramatizadas de asaltos a personas en una calle nocturna.

Para comprobar la validez en la detección de las amenazas del algoritmo propuesto en este artículo se presenta una comparación de su desempeño respecto a las tres propuestas [9]–[11] del estado del arte para la detección de acciones violentas. En [9] los autores introducen un descriptor basado en características binarias para codificar las regiones asociadas con los puntos de interés para la detección de eventos anormales. En [10] usan un clasificador SVM entrenado con características seleccionadas por un algoritmo Adaboost a partir de la información de magnitud y orientación del flujo óptico. Por su parte en [11] proponen un descriptor inspirado en un importante concepto de mecánica de fluidos para representar fuerzas a partir del flujo óptico y ser usado así en la detección de acciones violentas.

Adicionalmente, en el presente trabajo se realiza un conjunto de pruebas para comprobar experimentalmente en el algoritmo propuesto, la relación entre la capacidad de detectar una amenaza en forma temprana respecto a la generación de falsas alarmas.

El presente artículo está ordenado como se describe a continuación: en la sección II se realiza una descripción del algoritmo propuesto; el esquema de validación utilizado junto con el análisis de los resultados obtenidos se describen en la sección III; finalmente, en la sección IV se presentan las conclusiones y perspectivas de trabajos futuros.

II. Representación de Comportamientos Complejos Mediante Combinación de Eventos Base

Las ideas centrales en el método propuesto para la representación de comportamientos son la inclusión del historial de acciones de interés realizadas por los sujetos desde el momento que hacen su aparición en la escena y la expansión del análisis a más de una ventana de tiempo. Cada una de las acciones básicas de los sujetos y las interacciones entre individuos se consideran como parte integral de un mismo comportamiento y no de forma aislada.

Con este fin, se define un conjunto de M eventos base a partir de los cuales se constituyen los comportamientos complejos. Cada evento base E_m^i (m=1,2,...,M) generado por una persona $P^{(i)}$ tendrá un efecto sobre otra persona $P^{(j)}$, el cual depende del nivel de interacción existente entre los dos sujetos. Esto es representado mediante la matriz de interacción $A_m^{(i,j)}$ definida en (1). Cada vector fila de la matriz $A_m^{(i,j)}$ describe la relación de interacción entre $P^{(i)}$ y $P^{(j)}$ cuando se presenta el evento E_m^i realizado por $P^{(i)}$ en el tiempo discreto k.

$$A_m^{(i,j)}(k) = \left[1,\ \delta v_m^i(k),\ \delta t_m^{(i,j)}(k),\ \delta d_m^{(i,j)}(k)\right] \qquad (1)$$

El primer componente del vector $A_m^{(i,j)}(k)$ indica la ocurrencia de un evento tipo E_m^i en el tiempo k independientemente de las características particulares de dicho evento. El valor de este primer componente siempre será 1, indicando que el evento ha ocurrido. Mientras que para los tiempos cuando el evento E_m^i no ha ocurrido el vector representativo no se genera. La importancia de asignar el valor 1 a la ocurrencia del evento es que este valor será usado posteriormente por la función de consolidación (2), para establecer y ponderar el número de repeticiones en el tiempo de dicho tipo de evento. El componente δv_m^i cuantifica el cambio en la magnitud de la variable utilizada para detectar el evento. Su valor viene dado por $\delta v_m^i(t)=|\nu_m^i(t_N)-\nu_m^i(t_A)|$, donde ν_m^i es la variable utilizada para medir la acción que genera el evento E_m, (por ejemplo velocidad, angulo, etc.); t_A es el instante de tiempo antes de iniciarse la acción que genera el evento, mientras que t_N el instante de tiempo cuando se produce el evento. Los términos $\delta t_m^{(i,j)}$ y $\delta d_m^{(i,j)}$ en (1) ponderan el efecto sobre la distancia espacio-temporal que se produce entre los sujetos $P^{(i)}$ y $P^{(j)}$ como consecuencia del evento E_m^i. Estos componentes toman su valor a partir del procedimiento que se indica a continuación:

Se parte de las posiciones (x^i,y^i), (x^j,y^j) y los componentes de velocidad (v_x^i,v_y^i), (v_x^j,v_y^j) en un instante de tiempo t_k para los sujetos $P^{(i)}$ y $P^{(j)}$ respectivamente. Se define $d^{i,j}(t,t_k)$ como la distancia de separación entre los sujetos $P^{(i)}$ y $P^{(j)}$ en el instante de tiempo t estimada a partir de los datos de posición y velocidad dados en el tiempo t_k. El cuadrado de la distancia de separación $d^{i,j}(t,t_k)$ se obtiene de $k_1t^2+k_2t+k_3$, donde

$k_1=(v_x^i-v_x^j)^2+(v_y^i-v_y^j)^2$; $k_2=2(v_x^i-v_x^j)(x^i-x^j)+2(v_y^i-v_y^j)(y^i-y^j)$; $k_3=(x^i-x^j)^2+(x^i-x^j)^2$.

Se define $t_M^{(i,j)}(t_k)$ como el tiempo estimado para alcanzar la distancia mínima de separación entre $P^{(i)}$ y $P^{(j)}$ calculada a partir de la información espacio-temporal conocida en el tiempo t_k. Dicho tiempo viene dado por $t_M^{(i,j)}(t_k)=-\frac{k_2}{2k_1}$. En forma similar, se define $t_I^{(i,j)}(t_k)=\{t : d^{(i,j)}(t,t_k)=R_I\}$ como el tiempo estimado para que el sujeto $P^{(i)}$ se aproxime al sujeto $P^{(j)}$ dentro de un radio de interacción R_I. Para el presente trabajo se consideró un radio de dos metros (R_I=2m) como una distancia suficientemente cercana para una posible interacción entre las personas.

Finalmente, en la Tabla I se indica el valor asignado a los componentes $\delta t_m^{(i,j)}$ y $\delta d_m^{(i,j)}$ del vector representativo (1) asociado a un evento E_m^i. Donde, $t_{IA}=t_I^{(i,j)}(t_A)$; $t_{IN}=t_I^{(i,j)}(t_N)$; $t_{MA}=t_M^{(i,j)}(t_A)$ y $t_{MN}=t_M^{(i,j)}(t_N)$. El tiempo t_h es el tiempo para el cual una persona caminando a velocidad promedio sale del campo de visión de la escena. Este valor es obtenido para cada escenario en forma experimental. Para este fin, a partir de un

video de prueba se promedia el tiempo que le toma a varias personas caminando a velocidad normal atravesar la zona de interés en la escena. En el caso del escenario correspondiente al conjunto de datos PETS2014 este tiempo se estableció en $20s$, mientras que para el escenario del conjunto de datos UVS-dataset se estableció en $30s$.

En la primera columna de la Tabla I se establecen las condiciones lógicas que implican las respectivas asignaciones dadas en las columnas 2 y 3. Por ejemplo, en la segunda fila de la Tabla I, el término $C_1\bar{C}_2$ indica que la condición será verdadera si la variable de condición individual C_1 es falsa y la variable C_2 es verdadera. En este caso, los términos $\delta t_m^{(i,j)}$ y $\delta d_m^{(i,j)}$ en (1) toman los valores dados por $(t_h - t_{IN})$ y $(d^{(i,j)}(t_{IN},t_A) - R_I)$ respectivamente. La variable de condición individual c_1 viene dada por: $(0 \leq t_I^{(i,j)}(t_A) \leq t_h)$ la cual indica que C_1 será verdadera si a partir de los datos de posición y velocidad en el tiempo t_A se predice una aproximación entre los sujetos $P^{(i)}$ y $P^{(j)}$ dentro de una distancia de interacción R_I. De forma similar, se definen las restantes variables de condición como sigue: $C_2{=}0 \leq t_I^{(i,j)}(t_N) \leq t_h$; $C_3{=}0 \leq t_M^{(i,j)}(t_A) \leq t_h$ y $C_4{=}0 \leq t_M^{(i,j)}(t_N) \leq t_h$.

TABLA I
ASIGNACIÓN DE VALORES PARA LOS TÉRMINOS $\delta t_m^{(i,j)}$ Y $\delta d_m^{(i,j)}$

Condición	$\delta t_m^{(i,j)}$	$\delta d_m^{(i,j)}$
$C1C2$	t_{IA} - t_{IN}	0
$\bar{C}1C2$	t_h - t_{IN}	$d^{i,j}(t_{IN},t_A)$ - R_I
$C1\bar{C}2$	t_{IA} - t_h	R_I - $d^{i,j}(t_{IA},t_N)$
$\bar{C}1\bar{C}2C3C4$	0	$d^{i,j}(t_{MA},t_A)$ - $d^{i,j}(t_{MN},t_N)$
$\bar{C}1\bar{C}2C3\bar{C}4$	0	$d^{i,j}(t_{MA},t_A)$ - $d^{i,j}(t_{MA},t_N)$
$\bar{C}1\bar{C}2\bar{C}3C4$	0	$d^{i,j}(t_{MN},t_A)$ - $d^{i,j}(t_{MN},t_N)$
Otros casos	0	$d^{i,j}(t_h,t_A)$ - $d^{i,j}(t_h,t_N)$

Cada matriz de interacción $A_m^{(i,j)}$ descrita en (1) representa aisladamente las relaciones de interacción individuales entre $P^{(i)}$ y $P^{(j)}$ producto de un particular tipo de evento E_m^i, el cual realiza el sujeto $P^{(i)}$ en diferentes instantes de tiempo. Esta información es fusionada por medio de la función (2).

$$f_m^{(i,j)}(t) = U_m^{(i,j)}(t) A_m^{(i,j)}(t) I(d^{(i,j)}(t) < D_{Lim}) \quad (2)$$

La variable $U_m^{(i,j)}(t)$ corresponde a un vector fila $[1,1,...,1]$ de longitud N_m^i, igual al número de eventos E_m registrados para el sujeto $P^{(i)}$ desde su aparición en la escena hasta el tiempo t. Mención especial requiere el parámetro de distancia límite D_{Lim} en la función (2), mediante el cual se controla la sensibilidad entre la capacidad de anticipo en la detección de amenazas y la generación de falsas alarmas. La detección temprana de amenazas implica identificar posibles acciones criminales cuando los sospechosos se encuentran aún alejados de sus potenciales víctimas. Entre mayor sea el valor de la distancia que separa a los criminales de las víctimas en el momento de la detección de la amenaza, mayor será la capacidad de anticipación; sin embargo, también se incrementa la probabilidad de generar falsas alertas. La función indicadora $I(d_p^{(i,j)}(t) < D_{Lim})$ en (2) toma el valor 1 si la distancia de separación entre los sujetos $P^{(i)}$ y $P^{(j)}$ es menor a la distancia límite D_{Lim}. De esta forma, el valor dado a D_{lim} restringe la distancia de separación máxima a la cual se evaluará a una persona como posible sospechosa.

Con el fin de consolidar toda la información registrada durante la escena en un único vector representativo, entonces para cada sujeto $P^{(j)}$ se conforma el grupo de interacción G_I^j, con todos aquellos sujetos $P^{(i)}$ para los cuales se predice una potencial interacción de proximidad con $P^{(j)}$, de acuerdo con los respectivos datos de posición y velocidad en el tiempo actual t. Esto es, $G_I^{(j)}(t) = \left\{ P^{(i)} : 0 < t_I^{(i,j)}(t) < t_h \right\}$. Para cada $P^{(i)} \in G_I^{(j)}$ se define a su vez el grupo de interacción asociado $G_A^{(i)}(t) = \left\{ p^{(g)} : \tau_A^{(i,g)} > \tau_{min} \right\}$, el cual es conformado por el grupo de personas $P^{(g)}$ que previamente han interactuado con el sujeto $P^{(i)}$ durante un tiempo $\tau_A^{(i,g)}$ mayor que τ_{min}. En este trabajo se consideró $\tau_{min}{=}8s$; dado que este tiempo permite descartar como asociadas dos personas que pasan una junto a la otra sin interactuar. De esta forma, para cada sujeto $P^{(j)}$ en la escena, la función integradora $\Psi_m^{(j)}(t)$ definida en (3) reúne en un único vector resultante el registro de todos los eventos E_m realizados por todos los sujetos para los cuales se predice una interacción directa o indirecta con $P^{(j)}$.

$$\Psi_m^{(j)}(t) = \sum_{p^i \in G_I^j} f_m^{(i,j)}(t) + \sum_{p^i \in G_I^j} \sum_{p^g \in G_A^i} f_m^{(g,j)}(t) \quad (3)$$

En el caso de una actividad criminal, el primer término de la función (3) involucra los comportamientos de los sujetos sospechosos en forma individual, mientras que el segundo término refleja el comportamiento de un grupo de sospechosos trabajando en forma conjunta contra la víctima $P^{(j)}$.

Finalmente, dado un conjunto $\{E_1,E_2,...,E_M\}$ de eventos base de interés, el vector de características total $C^j(t)$, que representa el nivel de amenaza dado por el entorno para un sujeto $P^{(j)}$ en el instante de tiempo t, se obtiene de la concatenación de los vectores resultantes de (3) tal que $C^j(t) = \left\{ \Psi_1^j(t), \Psi_2^j(t), ..., \Psi_M^j(t) \right\}$. En este trabajo se tiene un conjunto de eventos base conformado por 5 tipos de eventos distintos $(M{=}5)$, cuya implementación es explicada en la siguiente sección. De esta forma, después del proceso de concatenación cada vector de características $C^j(t)$ tendrá un tamaño fijo de 20 componentes, los cuales después de un proceso de normalización son utilizados en un clasificador estándar SVM [20] que se encarga de la detección de los comportamientos sospechosos.

Una vez entrenado el clasificador se propone el algoritmo 1, el cual permite evaluar para un instante de tiempo determinado la presencia de una posible amenaza. El algoritmo recibe de la etapa de tracking la posición y los componentes de velocidad asociados a cada sujeto en la escena y entrega como salida para cada sujeto un valor que indica la detección de una potencial amenaza contra él.

Dado que la utilidad real del algoritmo es generar alertas tempranas durante su operación en línea, se requiere que su tiempo de respuesta sea lo suficientemente rápido para este propósito. El mayor nivel de complejidad computacional lo involucra la doble anidación en las líneas 6 y 9 lo cual sugiere una complejidad cuadrática $O(n^2)$, donde n corresponde al número de sujetos presentes en la escena en un instante de tiempo k. En realidad, en estos ciclos anidados el número

de operaciones está asociado a una combinatoria de tamaño $\frac{n!}{2!(n-2)!}$ siendo este número ostensiblemente menor al caso cuadrático. Considerando adicionalmente, que para este tipo de escenas no se espera un gran número de sujetos presentes y que las operaciones efectuadas son relativamente simples, el tiempo de ejecución del algoritmo no representa ninguna limitación. Para el caso específico de la implementación desarrollada se utilizó un computador con procesador I7 de 2400Mhz y el software Matlab, obteniéndose un tiempo promedio de ejecución para el algoritmo de $6ms$ sobre los dos conjuntos de datos de prueba. Esto no incluye el tiempo de la etapa de tracking por tratarse de un elemento externo al algoritmo propuesto.

III. Validación del Sistema

El conjunto de pruebas realizado tiene como objetivos: a) Comparar el desempeño en la detección de amenazas del algoritmo propuesto respecto a tres métodos del estado del arte. a) Evaluar la capacidad del esquema propuesto para generar alertas tempranas y comprobar la validez del método propuesto para ajustar el balance entre el nivel de anticipación a la amenaza y la generación de falsas alertas por medio del parámetro de distancia límite D_{Lim} definido en (2).

A. Conjunto de Datos de Validación

Existe un importante grupo de conjuntos de datos (datasets) disponibles públicamente destinados a evaluar algoritmos enfocados a la detección e interpretación de acciones y actividades humanas [21]. Algunos de estos datasets incluyen actividades tales como peleas y agresiones, que han sido ampliamente utilizados para detectar comportamientos violentos. La mayoría de estos datasets registran el momento cuando ocurre la acción violenta, sin embargo, dado que el principal interés del presente trabajo se centra en los comportamientos sospechosos antes de la agresión final, solamente resultan útiles aquellas secuencias de vídeo que incluyan comportamientos complejos previos. De acuerdo con las características de los conjuntos de prueba públicos consultados, el dataset PETS2014 es el que mejor se ajusta a este requerimiento, razón por la cual ha sido seleccionado para la fase experimental. Como complemento se utiliza el conjunto de datos propio al que se le denominó UVS-Dataset el cual contiene secuencias de video en infrarrojo de un escenario nocturno donde se representan acciones criminales de asalto.

PETS2014-Dataset: Es conformado por 22 secuencias de video cada una adquirida por cuatro cámaras alrededor de un camión estacionado. Se incluyen 3 secuencias de video que registran agresiones contra el conductor y 9 secuencias con potenciales amenazas contra el camión tales como: abrir la cabina, inspeccionar el interior, abrir las puertas o tocarlas. Adicionalmente incluye 10 secuencias con acciones habituales. En la Fig. 1 se muestra un ejemplo de las imágenes registradas.

UVS-Dataset: Recrea una calle en horas nocturnas, donde los criminales pueden tomar ventaja de las condiciones de baja iluminación para acechar y sorprender a sus víctimas. El dataset contiene 76 secuencias de video, grabadas con una cámara de vigilancia térmica Axis Q1910. La duración promedio de cada video es de 1.5 min. Las escenas representan acciones de asalto contra una víctima que incluyen un agresor (26 secuencias), dos agresores (17 secuencias) y tres agresores (7 secuencias). Adicionalmente se incluyen actividades habituales como caminar, cambiar de acera, caminar en grupo, aproximarse, encontrarse, saludar, hablar. La Fig. 2 muestra un ejemplo con imágenes del dataset.

Algoritmo 1: Detección de amenazas

Entradas:

k: Instante de tiempo discreto.

(x^i, y^i): Posición del sujeto $P^{(i)}$ en el instante de tiempo k, para $i = 1, \dots N$.

(v_x^i, v_y^i): Componentes de velocidad del sujeto $P^{(i)}$ en el instante de tiempo k, para $i = 1, \dots N$.

Salidas:

$D^i(k)$: Detección de potencial amenaza contra sujeto $P^{(i)}$ en el instante de tiempo k, para $i = 1, \dots N$.

Pasos:

Inicializar $D^i(k) = 0$ para $i = 1, \dots N$.

for all evento base E_m^i que se genere **do**

- Obtener la matriz de interacción $A_m^{(i,j)}$ definida en (1), para $j = 1, \dots, N;\ j \neq i$.
- Fusionar la información de cada matriz $A_m^{(i,j)}$ por medio de la función $f_m^{(i,j)}(k)$ definida en (2).

end for

for all sujeto $P^{(i)}$ **do**

- Obtener el grupo de interacción $G_I^{(j)}(t)=\left\{P^{(i)}:0<t_I^{(i,j)}(t)<t_h\right\}$

if $G_I^{(j)}(t) \neq \emptyset$ **then**

for all $P^i \in G_I^{(j)}$ **do**

- Obtener el conjunto $G_A^{(i)}(t)=\left\{P^{(g)}:\tau_A^{(i,g)}(t)>\tau_{min}\right\}$

end for

- Generar el vector representativo $\Psi_m^{(j)}(k)$ dado por la función integradora (3) para cada tipo de evento base E_m relativo al sujeto $P^{(j)}$.

end if

- Obtener el vector de características total $C^j(k)$ concatenando los vectores representativos de los M tipos de eventos base. $C^j(k)=\left\{\Psi_1^j(k),\Psi_2^j(k),...,\Psi_M^j(k)\right\}$ y normalizarlo.

if $\sum C^j(k) > 0$ **then**

- $D^j(k)$= Resultado al evaluar el clasificador SVM con $C^j(k)$.

end if

end for

Fig. 1. Imagen del escenario para el conjunto de prueba PETS2014 [19].

Fig. 2. Ejemplo con imágenes del conjunto de prueba UVS-dataset.

B. Eventos Base

Se define un conjunto de cinco eventos base que pueden constituir un comportamiento complejo. El evento E_1 que se encuentra asociado con una variación en la dirección de la trayectoria del sujeto $P^{(i)}$. El ángulo $\theta^i(k)$ de la trayectoria es estimado a partir del vector director $\bar{u} = (v_x^i, v_y^i)$. El rango de variación del ángulo es dividido en N_s sectores angulares, en este caso $N_s = 16$, con lo cual se logra simetría en los cuatro cuadrantes que conforman los 360^o y una resolución cercana a los 20^o que resulta suficiente para detectar cambios de dirección de interés.

El evento E_1 toma lugar si el ángulo de la trayectoria varía al menos α_1 sectores angulares respecto a la dirección previa. Para este trabajo se estableció $\alpha_1 = 2$ con el fin de reducir el efecto del ruido en la obtención del ángulo.

El evento E_2 se asocia con la variación en la velocidad de desplazamiento de la persona $P^{(i)}$. El rango de velocidades es dividido en N_I intervalos, en este caso con $N_I = 8$, valor con el que se logra un balance entre la detección de cambios de velocidad de interés y la reducción de falsos eventos debido al ruido. Un evento E_2 se produce cuando la velocidad de desplazamiento varía al menos α_2 intervalos respecto a la velocidad previa. Para este trabajo se estableció $\alpha_2 = 2$.

El evento E_3 toma lugar cuando la persona realiza una pausa mientras camina y se detiene en la escena durante α_3 segundos o más. Para este trabajo se estableció $\alpha_3 = 4$.

El evento E_4 es similar al anterior, pero en este caso la persona permanece estática en un lugar que no pertenece a una trayectoria habitual registrada en la escena. Esta clase de comportamientos podrían indicar una intención de permanecer oculto. Para establecer las trayectorias habituales, la escena es dividida en celdas de aproximadamente $1m^2$, posteriormente a partir de las rutas consideradas normales de todos los sujetos del conjunto de datos de entrenamiento se construye un histograma de tránsito. Para ello, cada que una persona pasa por una celda se incrementa una variable contadora asociada a dicha celda. El histograma de tránsito es almacenado y se utilizará entonces para detectar el evento E_4, el cual toma lugar cuando una persona se detiene en una celda vacía durante α_4 segundos o más. Para este trabajo se estableció $\alpha_4 = 4$.

El evento E_5 ocurre cuando la persona dobla su cuerpo con acciones tales como agacharse o sentarse. Para detectar este evento se utiliza la caja limitadora del cuerpo obtenida en la etapa de tracking. Para una explicación detallada del método de tracking utilizado referirse a [22] y [23]. El evento E_5 toma lugar cuando la altura de la caja limitadora cambia de tamaño en forma abrupta durante α_5 segundos o más. Para este trabajo se estableció $\alpha_5 = 5$.

C. Resultados Experimentales

De las 22 secuencias de video del conjunto de datos PETS2014 se obtuvieron un total de 19 víctimas, de las cuales 5 corresponden a personas agredidas y 14 son amenazas sobre el camión. Para el caso de las amenazas contra el camión, cada puerta del vehículo es procesada en forma equivalente a una persona, con la categoría de víctima por defecto, ubicada en las coordenadas espaciales de la puerta. Se obtuvieron adicionalmente 76 sujetos de la categoría normal (CN) cuyo número queda reducido a 33 después de la primera clasificación básica y que son utilizados en la fase de entrenamiento junto con los 19 de la categoría (CV). Por su parte, para el conjunto de datos UVS-Dataset se tiene en total 50 personas en la categoría víctimas (CV) y 229 en la categoría normal (CN). Los sujetos cuyo vector de característica esté conformado por valores 0's quedan automáticamente clasificados como (CN) y se excluyen para la etapa de entrenamiento. Después de este proceso inicial quedan finalmente 87 sujetos de la categoría (CN).

Para la validación se siguió el método estándar leave-one-out de la siguiente forma: se conforman aleatoriamente subconjuntos de datos de validación de 6 muestras (3 por categoría). Con las restantes muestras primero se realiza un proceso de sobre-muestreo para igualar el tamaño de las dos categorías y se entrena el clasificador SVM. El desempeño del clasificador se evalúa únicamente con los subconjuntos de datos de validación. Este proceso se repite hasta que todas las muestras hayan sido evaluadas como datos de validación.

Como indicador de desempeño se utiliza la curva ROC (Receiver Operating Characteristic) y su área bajo la curva, que permiten valorar el desempeño mostrado en la clasificación, así como el balance entre errores por fallas en la detección y errores por falsos positivos. En lo sucesivo de este artículo para hacer referencia al área bajo la curva ROC se utilizará el acrónimo AUC (Area Under Curve). Para obtener las curvas ROC primero se asigna un valor de probabilidad a cada muestra clasificada, para ello se usa la función de Platt [24] a la salida del clasificador. De esta forma, se obtiene un valor en el intervalo [0,1] para cada muestra evaluada. Seguidamente, mediante un procedimiento de umbralización variable se obtienen los valores de la curva ROC y el cálculo de su área.

1) Evaluación para la detección de amenazas: El primer grupo de pruebas tiene por propósito evaluar la capacidad del sistema propuesto para la detección de amenazas. Con este fin se compara su desempeño a partir de las curvas ROCs respecto a tres trabajos del estado del arte [9]–[11] para la detección de acciones de violencia. Tanto en [9] como en [10] no se requieren datos correspondientes a eventos sospechosos durante la fase de entrenamiento. Durante esta etapa el clasificador aprende los patrones considerados normales. Cuando un nuevo evento es evaluado, si sus patrones difieren significativamente de los patrones aprendidos será considerado un evento anómalo.

Para evaluar el método de Leyva [9] se utilizó el código disponible en [25]. Para el método de Gao [10] se utilizó el código disponible en [26]. En forma similar a [10] el tamaño de las celdas se definió de $4x4$, el número de rangos (bins) se fijó en 20 para el histograma asociado a la magnitud y 9 bins para el histograma asociado a la orientación del flujo óptico. El número de clasificadores débiles se estableció en 100. Para la implementación de Mohammadi [11] cada video fue aleatoriamente muestreado con bloques de tamaño $5x5x5$. A partir de los datos de entrenamiento se aprende un diccionario visual de tamaño $K = 100$, correspondientes a los centros de 100 clusters aprendidos usando el algoritmo *K-means*. Cada secuencia es dividida en sub-secuencias (clips) temporales de 15 frames de longitud y 5 frames de solapamiento. Cada clip es entonces descrito mediante un grupo de palabras (*bag-of-words*) a partir de los cuales se entrena un clasificador SVM de una clase, encargado de detectar clips con características de violencia.

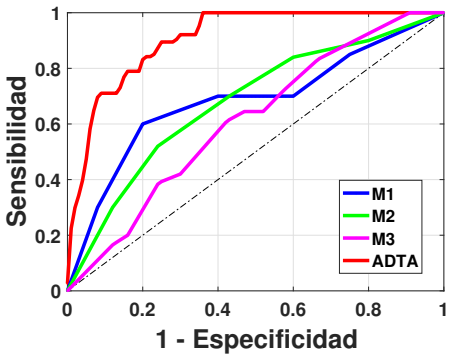
En la figura 3 se muestra las curvas ROCs correspondiente al desempeño del algoritmo ADTA propuesto junto con las curvas de los métodos de referencia a comparar. Para este grupo de pruebas se estableció el parámetro D_{Lim} en (2) con un valor de $10m$. Lo que significa que se podrán detectar sujetos sospechosos hasta un radio de $10m$ alrededor de la potencial víctima. En la Tabla II se reportan los resultados correspondientes al área bajo la curva AUC y el error EER (*Equal Error Rate*) de las respectivas curvas ROCs.

Los resultados muestran claramente un mejor desempeño del algoritmo ADTA propuesto respecto a las tres técnicas utilizadas para fines de comparación (M1, M2, M3). Algunas de las razones que explican esta mayor capacidad para la detección de amenazas son:

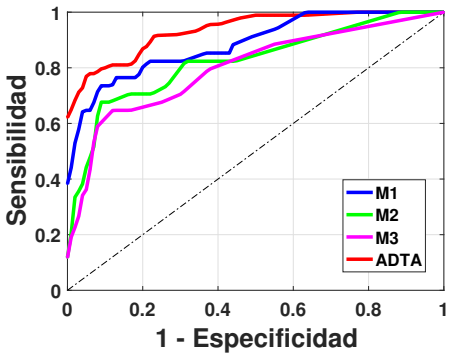
- Debido a que las técnicas M1, M2, M3 están basadas en características de bajo nivel son más susceptibles a cambios de iluminación y contraste en la escena, así como a las variaciones en la resolución de la imagen en la región de interés debidas a las diferentes distancias entre las personas y la cámara cuando ocurren los eventos.

- Las técnicas M1, M2, M3 solo pueden detectar acciones sospechosas que impliquen movimientos abruptos atípicos. Este tipo de movimientos suelen ser característicos en situaciones que involucran violencia física, pero no de comportamientos sospechosos donde no existe una interacción cercana entre los atacantes y las víctimas.

- A diferencia de las técnicas comparadas, el algoritmo ADTA propuesto permite establecer relaciones entre individuos sospechosos que trabajan en forma coordinada, aunque se encuentren en ubicaciones diferentes dentro de la escena.



(a) Curvas ROC - PETS-dataset



(b) Curvas ROC - UVS-dataset

Fig. 3. Curvas ROCs para la detección de agresiones. (a) Respuesta obtenida para el conjunto de prueba PETS2014. (b) Respuesta para el conjunto UVS-dataset. M1-(Gao [10]), M2-(Mohammadi [11]), M3-(Leyva [9]), ADTA-(algoritmo propuesto).

- Las técnicas comparadas solo utilizan la información de la ventana de tiempo durante la cual se desarrolla la agresión física, mientras que en el algoritmo propuesto se utiliza la información acumulada durante toda la escena.

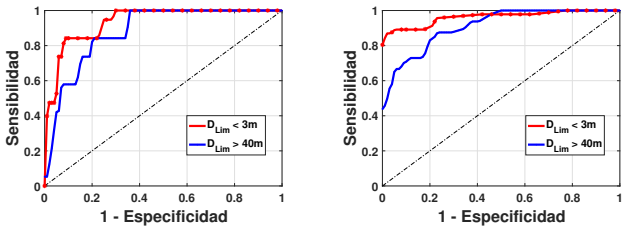
TABLA II
RESULTADOS DE LA CLASIFICACIÓN EN TÉRMINOS DE AUC Y EER

	PETS		UVS	
Method	AUC(%)	EER	AUC(%)	EER
M1- [10]	68.35	0.33	86.61	0.22
M2- [11]	67.80	0.36	82.25	0.26
M3- [9]	54.52	0.42	80.53	0.28
ADTA	**90.69**	**0.19**	**93.15**	**0.17**

2) Evaluación para la generación temprana de alertas: El propósito del siguiente grupo de pruebas es evaluar la capacidad del sistema propuesto para la generación temprana de alertas. Para este fin, se obtiene una medida del desempeño del algoritmo para detectar amenazas en función de la distancia de separación entre los sospechosos y las víctimas en el momento que se produce dicha detección. De esta forma, entre mayor sea la distancia de separación en el momento de generarse la alerta más preventiva será ésta. Como medida de desempeño se utiliza la curva ROC junto con su área bajo la curva. Para introducir variaciones en la distancia se utiliza

el parámetro D_{Lim} en (2), el cual permite que para un tiempo dado t se puedan detectar sujetos sospechosos hasta un radio de D_{Lim} alrededor de la posible víctima.

En la Fig. 4 se muestran las curvas de desempeño obtenidas sobre los conjuntos de datos PETS2014 y UVS-Dataset respectivamente. En cada figura el par de gráficas corresponden a las curvas ROC obtenidas para una distancia límite D_{Lim} menor de 3m y para una distancia límite mayor a $40m$. Las curvas muestran que al restringir la distancia de separación máxima entre las víctimas y los potenciales sospechosos se reduce el número de falsas alertas, sin embargo, en este último caso se pierde capacidad de anticipación en la detección de la amenaza.



(a) Curvas ROC - PETS-dataset (b) Curvas ROC - UVS-dataset

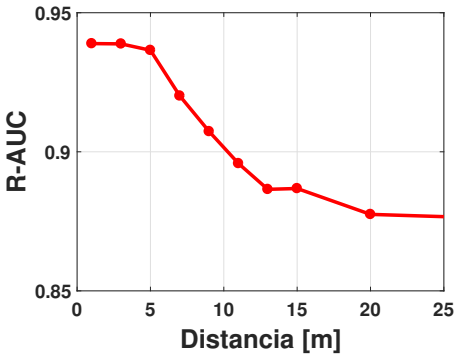
Fig. 4. Curvas ROC para valores extremos de la distancia límite de separación D_{Lim}. (a) Corresponden al conjunto de prueba PETS2014. (b) Corresponde al conjunto de prueba UVS-dataset.

Una generalización de los resultados anteriores se presenta en las Fig. 5(a) y 5(b), donde se muestra la variación del valor del AUC de las curvas ROCs obtenidas para diferentes valores de la distancia límite.

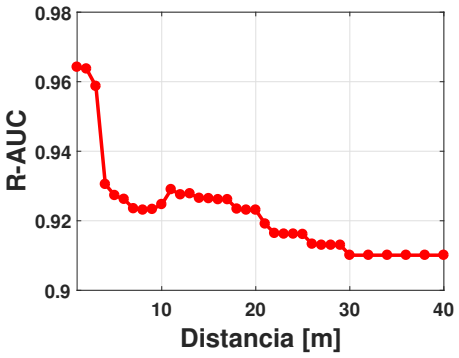
Para los dos datasets las curvas muestran un alto valor del AUC cuando la distancia de separación se limita a pocos metros. En el caso de PETS2014 con un AUC cercano al 0.94 para distancias menores a $5m$ y para UVS-dataset cercano a 0.96 para distancias menores de $3m$. Cuando se consideran distancias de separación mayores para la detección de potenciales amenazas, el valor del AUC disminuye como consecuencia de un incremento en la tasa de falsos positivos. Es de resaltar, sin embargo, que aún para distancias relativamente grandes (superiores a $30m$) de acuerdo con la geometría de la escena, se obtiene un desempeño alto con valores de AUC por encima de 0.87 para PETS2014 y 0.91 para UVS. Los resultados anteriores comprueban así la validez del esquema propuesto para la generación temprana de alarmas por posibles amenazas. Adicionalmente, se comprueba la utilidad del parámetro D_{Lim} como mecanismo de ajuste entre la capacidad de anticipo y la reducción de falsas alarmas.

IV. Conclusiones

A diferencia de trabajos previos enfocados en la detección de amenazas, donde el principal interés se centra en las acciones típicamente asociadas con violencia (peleas, golpes y ataques), el algoritmo propuesto en este trabajo explora el conjunto de acciones e interacciones que toman lugar antes que la actividad ilícita se presente. Con este enfoque se busca identificar comportamientos sospechosos que permitan alertar de una posible amenaza. Los comportamientos complejos son



(a) PETS-dataset



(b) UVS-dataset

Fig. 5. Variación del AUC respecto a la distancia límite de separación para la detección temprana de amenazas. (a) Respuesta obtenida para el conjunto de prueba PETS2014. (b) Respuesta para el conjunto UVS-dataset.

representados como una acumulación de eventos elementales conectados a partir de las interacciones presentes cuando dichos eventos toman lugar.

Una importante novedad que aporta el modelo propuesto es la introducción del parámetro de distancia límite D_{Lim} para el ajuste de sensibilidad entre la capacidad de anticipo y la generación de falsas alarmas. Los resultados sobre los dos conjuntos de datos de prueba mostraron que aún para los casos más exigentes de anticipación se obtienen valores de desempeño altos representados por un AUC alrededor de 0.9; lo cual comprueba un balance entre el acierto en la detección de amenazas y la reducción de falsas alarmas.

Al introducir nuevos elementos tales como un método eficiente y novedoso de representar en forma integral el conjunto de acciones de interés registradas durante la totalidad de la escena, el algoritmo propuesto constituye un aporte significativo a la línea de investigación en interpretación de comportamientos para aplicaciones de video vigilancia.

Los próximos trabajos se enfocarán a la reducción del número de falsas alarmas, para lo cual se implementarán y adicionarán modelos de interacción entre sujetos al esquema de representación en el algoritmo de detección temprana. Igualmente, con el ánimo de ampliar el rango de aplicabilidad del algoritmo propuesto, se realizarán pruebas sobre escenarios y actos delictivos diferentes a los utilizados en el presente trabajo.

AGRADECIMIENTOS

Agradecemos al grupo de investigación MIVIA de la Universidad de Salerno Italia por su apoyo en la construcción del conjunto de datos de prueba UVS-dataset. Igualmente, a la Universidad de Reading U.K por permitir el uso del conjunto de prueba PETS2014.

REFERENCIAS

[1] D. L. Cosmo, E. O. T. Salles, and P. M. Ciarelli, "Pedestrian detection utilizing gradient orientation histograms and color self similarities descriptors," *IEEE Latin America Transactions*, vol. 13, no. 7, pp. 2416–2422, July 2015.

[2] A. Magadan Salazar, I. Martin de Diego, C. Conde, and E. Cabello Pardos, "Evaluation of keypoint descriptors applied in the pedestrian detection in low quality images," *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1401–1407, March 2016.

[3] A. Santos Silva, F. Marcolino Quintao Severgnini, M. Lopes Oliveira, V. Matheus Santiago Mendes, and Z. M. Assis Peixoto, "Object tracking by color and active contour models segmentation," *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1488–1493, March 2016.

[4] D. L. Siqueira and A. Manso Correa Machado, "People detection and tracking in low frame-rate dynamic scenes," *IEEE Latin America Transactions*, vol. 14, no. 4, pp. 1966–1971, April 2016.

[5] D. Gonzalez Dondo, J. A. Redolfi, M. Griffa, G. M. Steiner, and L. R. Canali, "Target tracking system using multiple cameras and bayesian estimation," *IEEE Latin America Transactions*, vol. 14, no. 6, pp. 2713–2718, June 2016.

[6] R. Osorio, I. Lopez Juarez, M. Pena, V. Lomas, G. Lefranc, and J. Savage, "Surveillance system mobile object using segmentation algorithms," *IEEE Latin America Transactions*, vol. 13, no. 7, pp. 2441–2446, July 2015.

[7] J. M. Conejero, J. Hernandez, P. J. Clemente, R. Rodriguez Echeverria, J. C. Preciado, and F. Sanchez Figueroa, "Automatic configuration of video-surveillance applications: a model-driven experience," *IEEE Latin America Transactions*, vol. 13, no. 8, pp. 2700–2708, Aug 2015.

[8] T. Subetha and S. Chitrakala, "A survey on human activity recognition from videos," in *Proc. Int. Conf. Information Communication and Embedded Systems (ICICES)*, Feb. 2016, pp. 1–7.

[9] R. Leyva, V. Sanchez, and C. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, July 2017.

[10] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows." *Image Vision Comput.*, vol. 48, pp. 37–41, 2016.

[11] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative." in *AVSS*. IEEE Computer Society, 2015, pp. 1–6.

[12] M. R. Khokher, A. Bouzerdoum, and S. L. Phung, "Violent scene detection using a super descriptor tensor decomposition." in *DICTA*. IEEE, 2015, pp. 1–8.

[13] M. Ponti, T. S. Nazare, and J. Kittler, "Optical-flow features empirical mode decomposition for motion anomaly detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 1403–1407.

[14] A. Iscen, A. Armagan, and P. Duygulu, "What is usual in unusual videos? trajectory snippet histograms for discovering unusualness." *CoRR*, vol. abs/1401.0730, 2014.

[15] B. Najla and M. Emna Fendri, "Abnormal events detection based on trajectory clustering," *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, vol. 00, no. undefined, pp. 301–306, 2016.

[16] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection." *IEEE Trans. Circuits Syst. Video Techn.*, vol. 18, no. 11, pp. 1544–1554, 2008.

[17] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Bremond, "Towards abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, p. 1, 2016.

[18] L. Fuentes and S. A. Velastin, "Advanced surveillance: From tracking to event detection," *IEEE Latin America Transactions*, vol. 2, no. 3, pp. 206–211, Sep. 2004.

[19] P. J. Luis and F. J. M, "Pets 2014: Dataset and challenge," in *AVSS*. IEEE Computer Society, 2014, pp. 355–360.

[20] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199

[21] J. Chaquet, E. Carmona, and A. F. Caballero, "A survey of video datasets for human action and activity recognition." *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[22] D. Martinez, A. Saggese, M. Vento, H. Loaiza, and E. F. Caicedo, "Locally adapted gain control for reliable foreground detection." in *Lecture Notes in Computer Science*, G. Azzopardi and N. Petkov, Eds., vol. 9256. Springer, 2015, pp. 812–823.

[23] T.-S. Y. Jong-Min Jeong and J.-B. Park, "Kalman filter based multiple objects detection-tracking algorithm robust to occlusion," in *2014 Proceedings of the SICE Annual Conference (SICE)*, Sep. 2014, pp. 941–946.

[24] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 2000.

[25] https://cvrleyva.wordpress.com/.

[26] Piotr's computer vision matlab toolbox. [Online]. Available: http://vision.ucsd.edu/pdollar/toolbox/doc/index.html

Duber Martínez Torres, Ph.D., Ing. Electrónico (1997), Magister en Automática (2007) y Doctor en Ingeniería (2017) de la Universidad del Valle. Integrante del Grupo de Investigación en Percepción y Sistemas Inteligentes [PSI]. Sus áreas de interés son la visión artificial y el reconocimiento de patrones.

Humberto Loaiza Correa, Ph.D., Ing. Electricista (1990) y Magister en Automática (1993) de la Universidad del Valle; Doctor en Robótica y Visión Artificial (1999) de L'Université d'Evry, Francia. Profesor Titular y Director de la Escuela de Ingenierías Eléctrica y Electrónica de la Universidad del Valle. Codirector del Grupo de Investigación en Percepción y Sistemas Inteligentes [PSI]. Areas de interés son la robótica móvil, la visión artificial y el procesamiento digital de señales.

Eduardo Caicedo Bravo, Ph.D., Ing. Electricista (1984) y Magister en Automática (1995) de la Universidad del Valle; Doctor en Computación Industrial (1996) de la Universidad Politécnica de Madrid. Profesor Titular de la Escuela de Ingenierías Eléctrica y Electrónica de la Universidad del Valle. Director del Grupo de Investigación en Percepción y Sistemas Inteligentes [PSI]. Su interés de investigación actual son el campo de la robótica móvil, la inteligencia computacional y smart grids.