Evaluación de algoritmos de detección de eventos en sistemas de vídeo-seguridad

Informe Técnico

SemanticVideo.TR.2009.01

Julio 2009

Rafael Martin Nieto

Juan Carlos San Miguel

José M. Martínez

Índice del documento

1 Introducción

El proyecto *SemanticVideo* tiene como uno de sus principales objetivos la extracción de información semántica de la escena de vídeo captada. Para ello se propone investigar en el diseño y desarrollo de algoritmos de análisis de secuencias de vídeo para su aplicación en adaptación de contenido audiovisual. Uno de los campos de aplicación son los sistemas de vídeo-seguridad para los cuales se está investigando en diversos algoritmos de análisis de secuencias de vídeo. Uno de los problemas encontrados durante los trabajos ha sido la evaluación de los algoritmos. En el desarrollo de algoritmos, la forma de evaluar su eficiencia y robustez resulta complicada, ya que existe un campo muy amplio de investigación, y cada caso estudiado posee unas características que dificultan la creación de un sistema genérico. En los algoritmos para detectar eventos, a las dificultades ya existentes se debe añadir que su aparición es relativamente reciente, y aún no se ha trabajado suficientemente en su evaluación, lo que provoca una falta de referencias para basar las investigaciones.

En este documento se presenta un informe del trabajo relacionado con la evaluación de algoritmos y realizado dentro de la tarea T.1.3 "Mantenimiento, entorno de pruebas y difusión de resultados" del paquete 1 "Desarrollo del marco de trabajo".

La estructura de este informe es:

- La sección 1 presenta una breve visión general del estado del arte en sistemas de evaluación de algoritmos
- La sección 1 explica la metodología de evaluación de forma general
- La sección 4 explica un caso concreto de aplicación de la metodología de evaluación propuesta.
- Finalmente en la sección 5 se presentan las conclusiones del trabajo realizado.

2 Estado del arte

En [1] se explica que la evaluación de algoritmos de sistemas automáticos de vigilancia está limitada a secuencias cortas. Existe una falta de estandarización para comprobar la eficiencia de los algoritmos, y resulta difícil conocer las aplicaciones que pueden tener los algoritmos desarrollados en la vida real. En general, los resultados de la evaluación de los algoritmos dependen del *Content Set* (conjunto de prueba) elegido, lo que provoca que los algoritmos sean eficientes en un campo de aplicación muy reducido.

En [2] se comenta que la mayoría de las técnicas de evaluación descritas, no consideran la complejidad que presenta la metodología de evaluación. En toda evaluación existen unos parámetros de evaluación apropiados. Estos parámetros son característicos de cada caso particular, lo que dificulta la creación de un sistema de evaluación general.

En [3] los sistemas de evaluación se presentan como un aspecto muy importante en el desarrollo de aplicaciones. Con la evolución y mejora de la potencia de los ordenadores, se han podido crear sistemas que procesan secuencias en tiempo real. Con el desarrollo de estos sistemas, la evaluación toma gran importancia, ya que muestra las limitaciones y los campos de trabajo para los que están capacitados. Cada sistema de evaluación tiene su propio punto de vista, ya que se han diseñado en base a uno o varios sistemas concretos, es decir, el sistema de evaluación surge como consecuencia de la necesidad de evaluar un sistema de detección en secuencias determinadas.

En [4] el uso de sistemas de video vigilancia y el incremento de la demanda de estos sistemas, ha provocado un incremento en la cantidad de aplicaciones de detección automática. La robustez y fiabilidad de los algoritmos necesitan ser evaluados de alguna manera eficaz, y aquí es donde toma importancia la creación de sistemas de evaluación, ya que se necesita contrastar la funcionalidad teórica de los algoritmos con hechos experimentales.

En [5] se presentan unos objetivos similares a los de este informe, con la diferencia de que se estudia el seguimiento y detección de cara, texto, objetos y vehículos en vídeo. Para ello, se presenta una descripción formal de cada una de las tareas de evaluación y las directrices que se han seguido para crear las anotaciones.

En [6] se ha creado un conjunto de videos anotados, para que los usuarios puedan validar sus propios algoritmos y obtener resultados. En estos videos se pueden encontrar grabaciones multi-cámara, detección de objetos, seguimiento, reconocimiento, detección de eventos, etc.

En [7] se aportan varias razones por las que resulta interesante evaluar los algoritmos: conocer las mejoras durante el desarrollo, evaluar comparativamente con los competidores, medir el progreso en toda la comunidad científica, etc.

En [8] se estudian las limitaciones de las personas al vigilar de forma simultánea múltiples videos. En el estudio se demuestra que las personas son incapaces de detectar eficientemente los eventos que aparecen en los videos, sobre todo al aumentar el número de videos simultáneos.

En [9] se centran en uno de los principales problemas de este campo de estudio: los resultados de la evaluación de un algoritmo dependen en gran medida del conjunto de videos usado. Esto se debe a que cada algoritmo se diseña para funcionar en unas determinadas condiciones (interior/exterior, ciertas condiciones de iluminación, etc).

3 Metodología de evaluación

Si bien el objetivo principal de este trabajo era el desarrollo de un sistema de evaluación de algoritmos de detección de eventos, se ha querido trabajar desde un punto más amplio y riguroso. Por lo tanto, en primer lugar se han estudiado las diversas características y funcionalidades de los sistemas de evaluación deseados. Por lo tanto se ha desarrollado en primer lugar una metodología de evaluación de algoritmos de detección de eventos en sistemas de vídeo-seguridad (si bien es extensible a otros dominios de aplicación y tipos de algoritmos). Esta metodología, describe de forma general, los pasos y procesos que deberían existir en la mayoría de los sistemas de evaluación.

3.1 Selección del Content Set

Como punto de partida, con el fin de evaluar los distintos algoritmos se debe conseguir un conjunto de secuencias representativas, sobre los que se trabajará y se ejecutarán los algoritmos.

Para seleccionar los videos, es importante que sean de diferentes características, cubriendo los diversos casos que se puedan encontrar en la realidad. Este aspecto es importante ya que presentarán mayor o menor dificultad a los algoritmos para cumplir su función, lo que permitirá comprobar con mayor exactitud, la robustez y eficiencia de dichos algoritmos.

3.2 Clasificación del Content Set

Una vez conseguido el *Content Set*, se procede a identificar las características para clasificar los videos. El criterio de clasificación dependerá de los algoritmos a utilizar, y dependerá de características directamente relacionadas con el tipo de evaluación concreto.

El objetivo es crear diferentes categorías que presenten mayor o menor dificultad para la correcta funcionalidad del algoritmo, y así poder conocer la respuesta del algoritmo a los distintos problemas, pudiendo así solucionar cada aspecto por separado. De esta forma se podrá comprobar fácilmente las limitaciones de los algoritmos.

3.3 Anotación del Content Set: creación del Ground-Truth

Para comparar los resultados obtenidos por el algoritmo, se ha de definir previamente la solución real, es decir, lo que el algoritmo debería detectar en el mejor de los casos. Esta solución es lo que se denomina *Ground Truth*.

En el proceso de anotación, se comienza decidiendo los eventos que se van a anotar y que el algoritmo debería detectar. En este apartado es importante definir claramente el modo en el que se anota el evento (por ejemplo, *bounding box* + *frames* en los que aparece en evento), ya que ésta será la referencia con la que se compararán los resultados del análisis con el algoritmo.

A la hora de definir el modo de anotación hay que cubrir dos aspectos: la localización del evento y las reglas de anotación.

La localización del evento se refiere a los instantes y/o regiones del vídeo dónde ocurren los eventos, dando lugar a localizaciones en ***tiempo*** y en ***espacio***. Según el tipo de sistema bajo evaluación se tendrán en cuenta una, otra o ambas. Por lo tanto habrá que anotar tiempos (o *frames*) y/o regiones en función de la evaluación concreta. Posteriormente, a la hora de evaluar habrá que definir criterios para dar por válida una detección en función de ciertos márgenes de tolerancia. Resumiendo los aspectos de localización podemos tener los siguientes casos:

- ***Tiempo***. La evaluación se centra en la coincidencia de los tiempos anotados como *Ground Truth*, con los tiempos de los eventos detectados por el algoritmo. Se compara el solapamiento temporal entre la detección correcta y el resultado del algoritmo y se decide que la detección es correcta si el solapamiento supera un umbral mínimo.
- ***Espacio***. La evaluación se centra en la localización espacial de los eventos. En este caso se compara la descripción del evento, reflejada en el *Ground Truth*, con los resultados de detección del algoritmo, estableciendo unos rangos mínimos de coincidencia de las regiones anotadas y detectadas.
- ***Tiempo y Espacio***. La evaluación es una combinación de los dos puntos anteriores.

Con respecto a las reglas de anotación, se tiene que definir claramente cuando se considera que un evento ha ocurrido sin dar lugar a ambigüedades. Adicionalmente, también se tiene que considerar la duración del evento, es decir, aunque existan eventos que se puedan considerar 'instantaneos' el sistema de evaluación debe permitir cierta holgura temporal en la evaluación de la detección de dichos eventos.

Para ello, se deben definir las características y parámetros de cada evento, ya sean espaciales, temporales, o una mezcla de ambas, como se explicó en el apartado anterior.

Tras definir los eventos a detectar y las características y parámetros relevantes, se anotan todos los videos del *Content Set*, consiguiendo el *Ground Truth*.

3.4 Proceso de evaluación y presentación de resultados

Una vez seleccionado el *Content Set*, definidos los eventos, creado el *Ground-Truth* y ejecutado el algoritmo sobre los videos, se tiene todo lo necesario para comparar los resultados. Obviamente, los resultados del algoritmo deben obtenerse en un formato compatible con el *Ground-Truth*, para hacer más inmediata la comparación.

A la hora de dar los resultados, existen diversas medidas (e.g., detecciones correctas, falsos positivos, falsos negativos), entre las cuales destacan las conocidas como ***Precision*** y ***Recall***:

- ***Precision***. Cociente entre el número de eventos detectados correctamente y el número de candidatos que el algoritmo ha considerado como posibles eventos. Da una idea relativa de los falsos positivos que produciría el algoritmo.
- ***Recall.*** Cociente entre el número de eventos detectados correctamente y el número de eventos a detectar. Da una idea relativa de la capacidad del algoritmo para detectar los eventos.

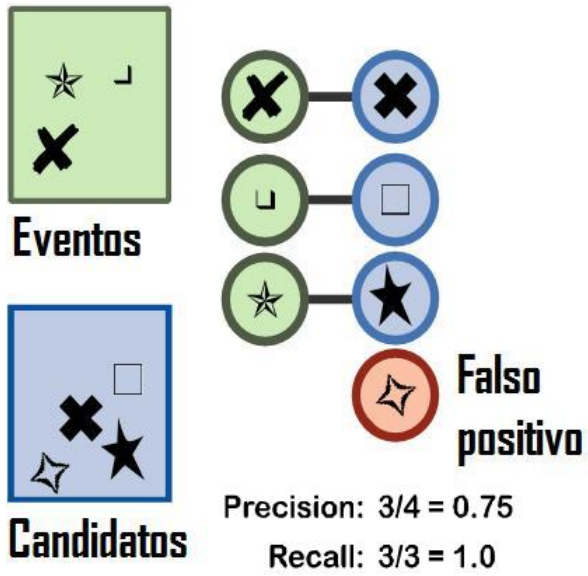

Figura 1: Ejemplo de *Precision* y *Recall*

4 Ejemplo de evaluación siguiendo la metodología propuesta

Tras definir la metodología de evaluación para los algoritmos, se presenta un caso concreto. Aplicando los pasos expuestos en la sección 1, se elegirá un determinado conjunto de videos en función del tipo de algoritmos que queremos evaluar (en este ejemplo serán eventos de interacción persona-objeto, concretamente: coger objeto, dejar objeto, abandonar objeto, robar objeto) y se clasificarán según sus características de complejidad.

Por otro lado se anota el Content Set para crear el Ground-Truth tras definir los eventos mediante su localización y parámetros. Adicionalmente se definen los rangos para considerar como detectado (o no) un evento, y tras la ejecución de los algoritmos se presentan los resultados.

Para este ejemplo, hemos usado dos algoritmos (descritos en [17]) de detección de los eventos arriba mencionados.

4.1 Selección del Content Set

Dado el tipo de eventos seleccionados, se escogió un *Content Set* con videos de video-vigilancia. Los videos se tomaron de los siguientes datasets públicos:

- *PETS2006 [10]*
- *PETS2007 [11]*
- *i-LIDS dataset for AVSS2007 [12]*
- *VISOR [13]*
- *CVSG [14]*
- *ITEA CANDELA project [15]*

4.2 Clasificación del Content Set

Los videos del *Content Set* se clasifican en distintas categorías según la complejidad de los mismos. Estas categorías estarán relacionadas con las tareas a realizar por los algoritmos a evaluar, y si bien pueden definirse algunas categorías genéricas (esto es, que aparecerán para diversos algoritmos), como la complejidad del fondo, los detalles concretos que definen a cada categoría variarán de algoritmo a algoritmo.

En nuestro caso, se clasificaron según dos aspectos relacionados con las tareas a realizar por los algoritmos de detección de eventos de interacción persona-objeto a evaluar:

- ***Dificultad para extraer los objetos del fondo:*** se define como la dificultad para extraer el movimiento y los objetos estacionarios en un escenario. Se relaciona, por ejemplo, con el número de los objetos, su velocidad, oclusión parcial o cambios de iluminación.
- ***Complejidad del fondo***: se define como la presencia de bordes, texturas múltiples y objetos pertenecientes al fondo, como árboles moviéndose, la superficie del agua, etc.

Los resultados de la clasificación se muestran en la tabla 1, mientras que la Figura 2 muestra ejemplos de secuencias de cada una de las categorías.

Categoría	Longitud	Complejidad	
		Extracción de fondo	Complejidad de fondo
C1	0h 1m 55s	Baja	Media
C2	0h 11m 11s	Baja	Alta
C3	1h 29m 18s	Media	Media
C4	2h 37m 21s	Media	Alta
C5	1h 26m 16s	Alta	Media

Tabla 1: *Categorías, longitudes y complejidad.*

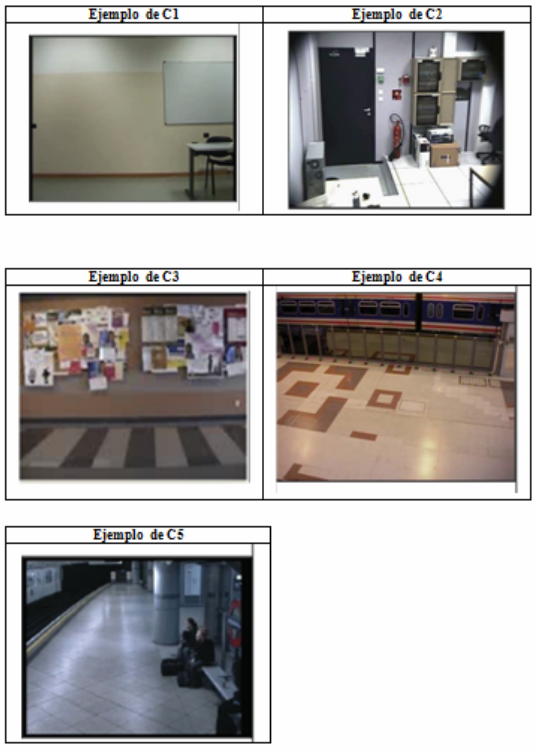

Figura 2: ***Ejemplos de secuencias de las distintas categorías***

4.3 Anotación del Content Set: creación del Ground-Truth

Para evaluar los algoritmos seleccionados de interacción persona-objeto, se ha decidido analizar cuatro posibles eventos:

- ***PutObject***: Un Sujeto se desprende de un objeto.
- ***AbandonedObject***: Tras desprenderse del objeto (esto es, producirse el evento *PutObject*), el sujeto se aleja del objeto, considerándose el objeto desatendido.
- ***GetObject*:** Un sujeto coge un objeto.

- ***StolenObject***: Tras coger un objeto (esto es, producirse el evento *GetObject*), el sujeto abandona con el objeto el lugar bajo vigilancia.

Tras observar los videos de las distintas categorías, se decidieron los siguientes criterios de anotación:

- ***PutObject:*** se anota el evento cuando el individuo se separa del objeto. Se tiene que ver la separación entre el objeto y el individuo.
- ***AbandonedObject:*** se anota el evento cuando tras un *PutObject*, el sujeto se aleja una distancia de aproximadamente 3 pasos o cuando el individuo sale de la imagen.
- ***GetObject***: se anota el evento cuando se aprecia claramente que un objeto existente deja de estar en su posición original. La persona puede estar enfrente o detrás del objeto. al cogerlo, por lo que en algunos casos no se anota hasta que la persona se aparta (momento en el que se "ve" el evento que puede haber ocurrido antes).
- ***StolenObject:*** se anota el evento cuando tras un *GetObject*, la persona que coge el objeto sale de la imagen con dicho objeto.

Para la creación del *Ground-Truth*, se ha utilizado el programa *Viper-GT* (http://viper-toolkit.sourceforge.net/products/gt/) de *Viper Annotation Toolkit*[16], anotándose las secuencias de vídeo *frame* a *frame* con los eventos definidos.

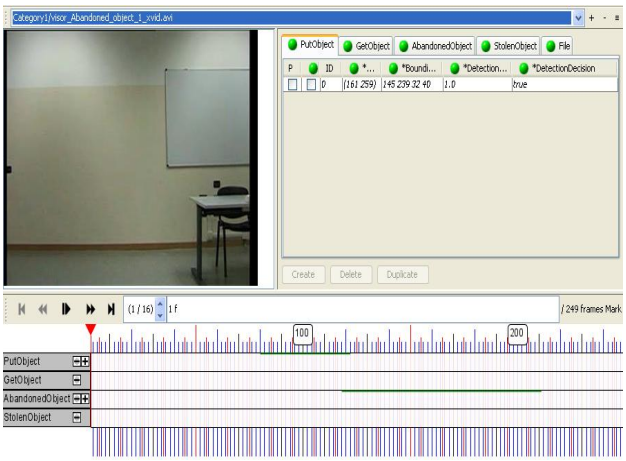

Figura 3: *Herramienta Viper-GT[16]*

Con este proceso, se obtuvo un fichero xml que contiene el Groud-Truth compuesto por las anotaciones de los eventos seleccionados, consistentes en la región (mediante una *bounding-box*) y un intervalo (consistente en un determinado número de f*rames* en las cuales "ocurre" el evento).

Respecto al intervalo de "ocurrencia" del evento, se decidió utilizar intervalos distintos para los eventos simples y complejos, debido a que los algoritmos son más lentos para estos últimos eventos. Se pueden diferenciar dos tipos de longitudes:

- ***Longitud Corta:*** Se corresponde con los frames en los que una persona que visualice el video podría considerar que se produce el evento.

- ***Longitud Larga:*** Se corresponde con los frames de la longitud corta, añadiendo a continuación un intervalo de frames en los que podría aparecer el evento detectado, debido al retardo computacional en el proceso de análisis del algoritmo.

Por ello, se tomó el criterio de longitudes mostrado en la tabla 2.

Eventos	Longitud Corta(Frames)	Longitud Larga (Frames)
Sencillos	50	300
Complejos	100	500

Tabla 2: *Longitudes de eventos.*

4.4 Proceso de evaluación y presentación de resultados

Para comparar los resultados de los algoritmos seleccionados (descritos en [17]) con el Ground-Truth, se utilizó la herramienta *Viper-PE* (http://viper-toolkit.sourceforge.net/products/gt/) de *Viper Annotation Toolkit*[16]. Esta herramienta necesita dos ficheros, uno de propiedades y otro de configuración. A continuación se describen algunos de los parámetros de dichos ficheros:

- Fichero "config.epf":
  - BEGIN/END_OBJECT_EVALUATION. eventos que se pretenden evaluar y atributos.
  - BEGIN/END_RESULT_FILTER. Eventos que queremos filtrar del fichero de resultados.
  - BEGIN/END_GROUND_FILTER. Eventos que queremos filtrar del fichero de ground truth.
- Fichero "properties.pr":
  - *Level*. Tipo de análisis. Existen cuatro tipos: detección, localización, comparación estadística y *matching* de objetivos. Nosotros nos centraremos en inspeccionar las 2 primeras opciones.
  - *Range_metric*. Es la métrica que produce los resultados.
  - *Range_tol*. Es la tolerancia temporal máxima (en frames).
  - *XXXX_tol*. Son las tolerancias de los atributos definidos en los eventos. En principio estos valores deberían ser lo más alto posibles para evitar que perdamos eventos. Al existir un mínimo de coincidencia, aunque se aumenten las tolerancias no se aumentarán los falsos positivos.

Al existir distintos grados de dificultad se consideró interesante analizar por separado la detección de los dos eventos sencillos y la detección de todos los eventos. En las tablas 3 y 4, se pueden observar los resultados obtenidos.

Eventos Sencillos				
Categorías	Longitud Corta		Longitud Larga	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
C1	100	68.96	100	68.96
C2	36.11	22.03	41.66	25.42
C3	34.00	35.36	37.38	38.87
C4	20.38	7.16	25.24	8.87
C5	30.68	1.08	43.18	1.52

Tabla 3: *Resultados de análisis en los 2 eventos sencillos.*

Resultados Totales				
Categorías	Longitud Corta		Longitud Larga	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
C1	52.63	66.66	52.63	66.66
C2	24.61	17.02	27.69	19.14
C3	31.96	36.42	34.32	39.12
C4	20.92	8.97	25	10.72
C5	37.67	1.68	47.26	2.11

Tabla 4: *Resultados de análisis en los 4 eventos considerados.*

Como se puede observar, los algoritmos probados funcionan bien para los eventos sencillos sobre secuencias de video sencillas. Al aumentar la complejidad de los videos, los resultados empeoran notablemente, destacando especialmente los problemas que presenta el algoritmo al analizar secuencias con mayor complejidad de fondo (como era de esperar). En cualquier caso en este informe estamos presentando la metodología de evaluación, sin entrar a evaluar en si los algoritmos concretos.

5 Conclusiones

En este informe, se ha presentado una metodología de evaluación de algoritmos de detección de eventos en sistemas de vídeo-seguridad. Esta metodología presenta una posible solución a la falta de estandarización de los métodos para evaluar algoritmos. Se presenta además un ejemplo de uso de evaluación siguiendo dicha metodología.

Adicionalmente a la evaluación comparativa de algoritmos, la evaluación de un único algoritmo permite diseñar mejoras en el mismo: al obtenerse resultados cuantitativos de la eficacia del algoritmo se pueden conocer los problemas y las dificultades concretas frente a las que el algoritmo no funciona, y se puede desarrollar y modificar el algoritmo sabiendo, con un cierto margen, que aspectos se deben mejorar o modificar.

Referencias

[1] L.M. Brown, A.W. Senior, Y.-L. Tian, J. Connell, A. Hampapur, C.-F. Shu, H. Merkl, M. Lu, “Performance Evaluation of Surveillance Systems Under Varying Conditions”, Proc. of IEEE Performance Evaluation of Tracking and Surveillance Workshop, 2005.

[2] N. Lazarevic-McManus, J. Renno, G.A. Jones, “Performance Evaluation in Visual Surveillance using the F-Measure”, UK KT1 2EE, 2006.

[3] A. T. Nghiem, F. Bremond, M. Thonnat, V. Valentin, “ETISEO, performance evaluation for video surveillance systems”, Proc. of IEEE Int. Conf. of Advanced Video-based Signal Surveillance, 2007.

[4] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H.S. Loos, M. Merkel, W. Niem, J.K. Warzelhan, J. Yu, “A Review and Comparison of Measures for Automatic Video Surveillance Systems”, EURASIP Journal on Image and Video Processing, 2008.

[5] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, J. Zhang, ramework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):319-336, Feb. 2009.

[6] R. Vezzani, R. Cucchiara, "Annotation Collection and Online Performance Evaluation for Video Surveillance: the ViSOR Project", Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS2008), 2008.

[7] D. Roth, E. Koller-Meier, D. Rowe, T.B. Moeslund, L. Van Gool, "Event-Based Tracking Evaluation Metric," Proc. of IEEE Workshop on Motion and Video Computing (WMVC 2008), 2008

[8] N. Sulman, T. Sanocki, D. Goldgof, R. Kasturi, "How effective is human video surveillance performance?," Proc. of . International Conference on Pattern Recognition (ICPR 2008), 2008.

[9] A.T. Nghiem, F. Bremond, M. Thonnat, R. Ma, “A New Evaluation Approach for Video Processing Algorithms” Proc. of IEEE Workshop on Motion and Video Computing (WMVC 2007), 2007.

[10] PETS2006: Ninth IEEE International Workshop on Performance Evaluations of Tracking and Surveillance, June 2006. URL http://pets2006.net/

[11] PETS2007: Tenth IEEE International Workshop on Performance Evaluations of Tracking and Surveillance, October 2007. URL http://pets2007.net/.

[12] i-LIDS dataset for AVSS2007: Fourth IEEE International Conference on Advanced Video and Signal based Surveillance, September 2007. URL:http://www.elec.qmul.ac.uk/sta_nfo/andrea/avss2007_d.html

[13] VISOR: Video Surveillance Online Repository. URL: http://imagelab.ing.unimore.it/visor

[14] CVSG: A Chroma-based Video Segmentation Ground-truth. URL: http://www-vpu.ii.uam.es/CVSG/

[15] ITEA CANDELA project: scenarios for abandoned object detection. URL: http://www.multitel.be/~va/candela/abandon.htmlCategorization

[16] D. Doermann, D. Mihalcik, "Tools and techniques for video performances evaluation", Proc. of IEEE Int. Conference on Pattern Recognition (ICPR 2000), 2000.

[17] J.C. San Miguel “Ontology-guided semantic video description using feedback between signal processing stages”, Master thesis (Advisor: Jose M. Martinez), Universidad Autónoma de Madrid, 2008.