Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 152-161

Tipo de artículo: Artículo Original
Temática: Reconocimiento de patrones
Recibido: 30/09/2015 | Aceptado: 20/12/2015

Reconocimiento de armas en imágenes de rayos X mediante Saco de Palabras Visuales

Weapons recognition in X-ray images using Bag of Visual Words

David Castro Piñol[1*], Frank Sanabria Macias[1], Enrique Marañón Reyes[1], Felipe Rodriguez Arias[1]

[1]Centro de Estudios de Neurociencias, Procesamiento de Imágenes y Señales (CENPIS). Universidad de Oriente, Cuba

[*]Autor para correspondencia: davidpinol@uo.edu.cu

Resumen
El diseño de un sistema automático que reconozca objetos peligrosos en imágenes de rayos X de equipos de inspección ha sido un problema complejo en los últimos años. La inspección de equipajes por rayos X presenta limitantes en cuanto a la eficiencia en el reconocimiento de objetos peligrosos y la demora que se toma el proceso. No existe una herramienta software que detecte automáticamente la presencia de armas en imágenes de rayos X y facilite el trabajo del operador de inspección. En este trabajo se desarrolló e implementó un algoritmo para el reconocimiento de armas cortas en imágenes de rayos X usando el método Saco de Palabras Visuales. Para realizar esto se implementó una etapa de pre-procesado, se construyó el vocabulario de palabras visuales que tuviera el mejor comportamiento frente a este tipo de imágenes, se representó un conjunto de imágenes mediante los histogramas de palabras visuales y se realizó el entrenamiento de un clasificador de tipo Máquina de Soporte Vectorial. Este algoritmo se desarrolló sobre la plataforma Matlab y con el apoyo de la biblioteca de funciones VLFeat. Se realizaron diversos experimentos variando los parámetros del método obteniéndose como mejor resultado una razón de verdaderos positivos de un 97.12 % y una razón de falsos positivos de 7.4 %. Estos resultados muestran que el algoritmo implementado puede servir de apoyo al personal de inspección, aumentar la rapidez del proceso y mejorar la eficiencia en el reconocimiento de armas en las imágenes de rayos X del sistema de inspección de equipajes.

Palabras claves: Saco de Palabras Visuales, Máquina de Soporte Vectorial, imágenes de rayos X

Abstract
An automatic system's design that recognizes dangerous objects in baggage X-ray images has been a complex problem in recent years. X-ray inspection has difficulties because of the low efficiency in automatic recognition of dangerous objects and inspection process delay. It doesn't exist a software application that automatically detects weapons in those images and reduce the workload of screeners. In this project was developed and implemented an algorithm for recognizing handguns in X-ray images using the Bag of Visual Words method. In order to achieve this, it was implemented a preprocess, was built a vocabulary of visual words with the better performance for this kind of images, it was represented a set of images by histograms of visual words and it was trained a Support Vector Machine classifier. This algorithm was developed in Matlab platform using VLFeat library. It was performed several experiments handling tunable parameters, getting the most relevant result a true positive

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 152-161

rate of 97.12 % and a false positive rate of 7.4 %. These results show that the implemented algorithm could be a support for inspection screeners and hence increase inspection speed and increase the efficiency of weapons recognition in X-ray images of inspection system.

Keywords: *Bag of Visual Words, Suport Vector Machines, X-ray images*

Introducción

Las imágenes de rayos X constituyen una importante tecnología para aplicaciones de seguridad en los equipos de inspección presentes en puertos y aeropuertos. A pesar de su alta efectividad, los sistemas de inspección actuales con esta tecnología tienen algunas dificultades. Las mismas están relacionadas principalmente con la posibilidad que el personal que opera el sistema cometa errores, ya sea por agotamiento visual, falta de un entrenamiento correcto, poca experiencia, etc. Esta situación hace que dichos sistemas, hasta el momento, activen muchas alarmas cuando no hay objetos peligrosos, haciendo más lento el proceso de inspección o situación más peligrosa aún, se deje pasar un objeto peligroso. De manera que se hace necesario el diseño de un sistema semiautomático del proceso de inspección para reducir la carga de trabajo, mejorar la eficiencia de la clasificación y aumentar la velocidad de inspección, (BAŞTAN et al., 2011). Se habla de un sistema semiautomático porque el objetivo no es desplazar al personal entrenado sino fortalecer y complementar su trabajo.

Los equipos de rayos X de energía dual forman la imagen enviando dos rayos de energías diferentes. A partir de la atenuación del rayo recibido, en cada posición (píxel), se estima la densidad y el número atómico efectivo de los materiales. En las imágenes formadas, el tono del color va a estar relacionado con el número atómico efectivo del material. El naranja se usa para materiales orgánicos, el azul para materiales metálicos y verde para materiales intermedios. Por su principio de formación, las imágenes de rayos X se caracterizan por presentar objetos solapados y no ocluidos como en las imágenes del espectro visible. Además pueden resultar ruidosas debido a la baja energía de los rayos emitidos por el equipo y pueden encontrarse objetos en diversos puntos de vista. De esta manera el reconocimiento de objetos en dichas imágenes se torna un problema complejo (BAŞTAN et al., 2011).

En los últimos años se han realizado investigaciones de algoritmos de visión por computadora para aplicarlos a las imágenes de rayos X que entregan los equipos de inspección. Uno de los métodos que ha tenido muy buenos resultados es Saco de Palabras Visuales o *Bag-of-Visual-Words* (BoVW) propuesto por (CSURKA et al., 2004) para búsqueda de imágenes por contenido y clasificación de objetos en imágenes de espectro visible. Los trabajos realizados por Başţan y Turcsany (BAŞTAN et al., 2011, 2013; TURCSANY et al., 2013) aplican el método BoVW en el contexto de imágenes de rayos X de equipos de una sola vista para reconocer objetos peligrosos. Sin embargo existe una bibliografía limitada sobre el tema y las investigaciones realizadas

que usan BoVW presentan ciertas diferencias en la experimentación como son las bases de datos utilizadas, la cantidades de imágenes entre otros. Además no se presentan los efectos de parámetros importantes en la clasificación como el kernel. El presente trabajo tiene como objetivo la realización de una exploración más intensiva de BoVW en imágenes de rayos X, para desarrollar un algoritmo de reconocimiento de armas cortas con mejor desempeño.

El algoritmo de clasificación basado en BoVW para el reconocimiento de armas que se propone se utilizaría sobre una ventana deslizante, (JONES and VIOLA, 2001) para la detección de estos objetos en las imágenes de rayos X. En el proceso de entrenamiento del algoritmo clasificador no se utilizaron las imágenes originales, sino las instancias de la ventana deslizante.

Método Saco de Palabras Visuales

El método Saco de Palabras Visuales está constituido por dos fases. La primera se basa en la construcción del vocabulario de palabras visuales de una base de datos de imágenes. Para ello, se extraen rasgos visuales de todas las imágenes con algún metodo de extracción de características. Luego se utiliza un algoritmo de agrupamiento (*clustering*), usualmente k-means, que crea grupos de rasgos visuales similares entre sí. Los centros de cada grupo son llamados palabras visuales.

La segunda fase es la representación de la imagen a clasificar mediante un histograma de palabras visuales. Este histograma se construye mediante una cuantificación vectorial de los rasgos extraídos de la imagen con el vocabulario de palabras visuales. Se trata de asignar cada rasgo a la palabra visual más cercana.

Con los histogramas de palabras visuales se puede construir un clasificador binario. La clasificación es binaria entre objetos peligrosos y otros objetos. El clasificador más utilizado con BoVW es la Máquina de Soporte Vectorial o *Support Vector Machine* (SVM), (VAPNIK, 1998). El entrenamiento de la SVM consiste en encontrar el hiperplano que maximice el margen de separación entre las dos clases.

Para poder tratar con datos que no son linealmente separables se introduce la función kernel y la función de pérdida. El kernel permite expandir los datos a un espacio de mayor dimensión buscando una mejor separación lineal entre las clases. Existen diferentes tipos de kernels. Los kernels lineales son más eficientes en el entrenamiento pero los no lineales logran una mejor separación de clases,(CHATFIELD et al., 2011), no obstante los kernels homogeneos aditivos agrupan ambas ventajas,(VEDALDI and ZISSERMAN, 2012). La función de pérdida permite manejar muestras mal clasificadas en los datos de entrenamiento y da una medida del error cometido.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 152-161

Materiales y métodos

En esta sección se presenta la implementación del método BoVW en el contexto de imágenes de rayos X que permite el reconocimiento de armas cortas. La misma fue desarrollada sobre la plataforma MATLAB 2014 y la biblioteca de funciones VLFeat 0.9.18, (VEDALDI and FULKERSON, 2008; VEDALDI and ZISSERMAN, 2011).

Pre-procesamiento

Las armas de fuego, en general, están compuestas por partes metálicas. Los metales en las imágenes de rayos X presentan colores en diferentes tonos de azul. De manera que constituye una ventaja para la tarea de clasificación si se logran extraer únicamente los rasgos de color azul. Se ha demostrado que esta estrategia ha tenido buenos resultados (BAŞTAN et al., 2011). Se procede a explicar la propuesta de una etapa de pre-procesado.

Primero se realiza una segmentación de regiones de color azul mediante el método de la esfera (*sphere*), (GONZALEZ and WOODS, 2002), obteniéndose una imagen binaria. Después se realizó una operación morfológica de cierre para rellenar huecos, un filtrado por áreas para rechazar zonas muy pequeñas y una operación morfológica de dilatación para incluir posibles partes no metálicas presentes en las armas. Se puede apreciar en la figura 1 las regiones a las que se le extraen las características con el algoritmo PHOW. Este paso resulta conveniente para la aplicación del clasificador en un esquema de ventana deslizante, eliminando todas las ventanas que no contengan áreas significativas con componentes metálicos.

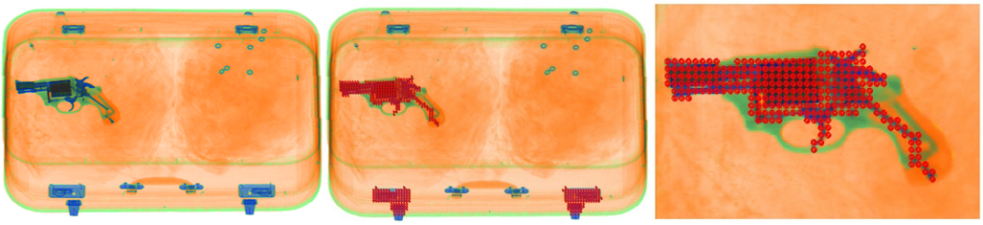

Figura 1. Características extraídas después del pre-procesado

Extracción de características

En este trabajo se utilizó para la extracción de características el algoritmo PHOW (*Pyramid Histogram Of Visual Words*) propuesto por Bosch (BOSCH et al., 2007). PHOW realiza un muestreo denso de puntos con espacio de M píxeles (se seleccionó M=4) a cuatro escalas fijas (definidas a priori) y está basado en los descriptores de SIFT, (LOWE, 2004). Las escalas son definidas modificando el ancho de la ranura espacial del descriptor SIFT a 4, 6, 8 y 10 píxeles respectivamente. Los características fueron extraidas de cada uno de los

canales HSV. El muestreo denso es conveniente debido a que los objetos en imágenes de rayos X poseen poca textura y es necesario obtener más información de ellos para su clasificación.

Construcción de vocabularios de palabras visuales

El algoritmo utilizado para el agrupamiento de características fue el clásico k-means. En este trabajo se construyeron vocabularios de tamaños 1000, 3000 y 5000 palabras, utilizando todas las imágenes originales de la base de datos. Se construyeron dos vocabularios por cada tamaño, el vocabulario universal y el vocabulario metálico. El vocabulario universal se construyó con las características extraídas sobre toda la imagen. Para el vocabulario metálico se utilizaron las características de las regiones que quedaron después de realizar el preprocesado. Se espera una mejor representación de las imágenes con un vocabulario construido con las regiones metálicas de los objetos de interés según se infiere de un razonamiento de Perronnin (PERRONNIN et al., 2006).

Histogramas de palabras visuales y entrenamiento de la SVM

Los histogramas se construyeron usando asignación dura (*hard assigment*) además de usar el factor de normalización. Los histogramas de palabras visuales van a tener implícitamente cierta información espacial, (YANG et al., 2007) debido al solapamiento de las características visuales extraídas con PHOW. Esta situación es ventajosa ya que adiciona información de las relaciones espaciales entre rasgos. En este trabajo se utilizaron los siguientes kernels homogéneos aditivos de VLFeat: Intersección, χ^2 y Jensen-Shannon. Además se implementó el Hellinger. Se utilizaron las funciones de pérdida: *Hinge*, *Square hinge* (hinge2), *Square* (L2), *Linear* (L1) y *Logistic* presentes en VLFeat. El parámetro de regularización λ del algoritmo de entrenamiento se determinó empíricamente con valor de $\lambda = 0{,}0001$.

Resultados y discusión

En esta sección se presentan los resultados obtenidos con varios vocabularios y clasificadores binarios SVM. Para el proceso de evaluación se utilizó validación cruzada de 3 segmentos, presentando el promedio de las curvas ROC (*Receiver Operating Characteristict*). El área bajo la curva (AUC de sus siglas en inglés) se utilizó como métrica para seleccionar la mejor configuración en cada caso. Por las exigencias de la aplicación se priorizó la razón de verdaderos positivos (TPR de sus siglas en inglés) sobre la razón de falos positivos (FPR de sus siglas en inglés) es decir la elección del punto de operación debe estar por encima de la diagonal principal del espacio ROC.

Descripción de la base de datos

La base de datos fue conformada por imágenes de equipos de inspección con rayos X de energía dual. Tiene un total de 948 ficheros de imágenes en formato PNG, con resoluciones que varían alrededor de 1000x600 píxeles.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 152-161

Para la validación cruzada se prepararon un conjunto de 312 imágenes positivas y 567 negativas. Las imágenes positivas fueron ventanas que pertenecen a la clase arma corta. Las imágenes negativas fueron ventanas, escogidas de manera aleatoria, de objetos metálicos (y/o parte de ellos) que no son armas. Las armas en la base de datos aparecen en diversas condiciones con diferentes niveles de complejidad para el reconocimiento. Estas situaciones fueron catalogadas en los siguientes seis grupos, 1: objetos en oclusión propia (123 armas), 2 Solapadas con objetos metálicos (94 armas), 3 Partes no metálicas (30 armas), 4 Distorsión geométrica de la adquisición (15 armas), 5 Parcialmente desarmadas (12 armas) y 6 Vista frontal simple (92 armas). También hay armas que tienen una combinación de estas situaciones. En la figura 2 se presentan una muestra de cada grupo.

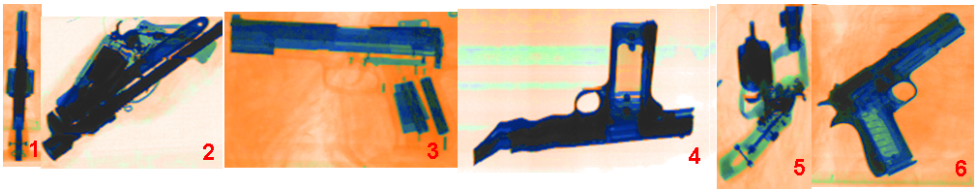

Figura 2. Situaciones de las imágenes positivas utilizadas

Evaluación del clasificador

En la figura 3 a) se observan las curvas ROC de los cuatro kernels homogéneos. Los demás parámetros del entrenamiento se mantienen con su configuración inicial. Los kernels χ^2 y Hellinger presentaron mejor comportamiento, siendo superior el χ^2 que será utilizado en los siguientes experimentos. En la figura 3 b) aparecen las curvas ROC para las funciones de pérdida. Las funciones de pérdida con mayor AUC son L1 y L2. Aunque la diferencia no es significativa, se decidió seleccionar como función de pérdida a L2 debido a que es menos costosa computacionalmente que L1, (VEDALDI and FULKERSON, 2008). En la figura 3 c) aparecen las curvas ROC de cada vocabulario construido. Se puede observar como el AUC de las curvas pertenecientes a los vocabularios metálicos son superiores que la de los vocabularios universales y a medida que aumenta el tamaño del vocabulario es superior la diferencia en cada par de vocabularios con igual tamaño. Por encima de la diagonal principal se destacan los vocabularios: 1000 universal, 1000 metálico y 5000 metálico. Se seleccionó el punto de operación con TPR=97.12 % y FPR=7.4 % que aparece señalado en la figura 3 c). Este punto pertenece simultáneamente a las curvas del vocabulario metálico de 1000 y 5000 palabras visuales respectivamente. Es preferible utilizar el vocabulario metálico de tamaño 1000 frente al de 5000 por razones de costo computacional, de manera que este es el vocabulario propuesto. El punto de operación seleccionado tiene una precisión de PPV= 87.57 % y una exactitud de ACC=94.2 %.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 152-161

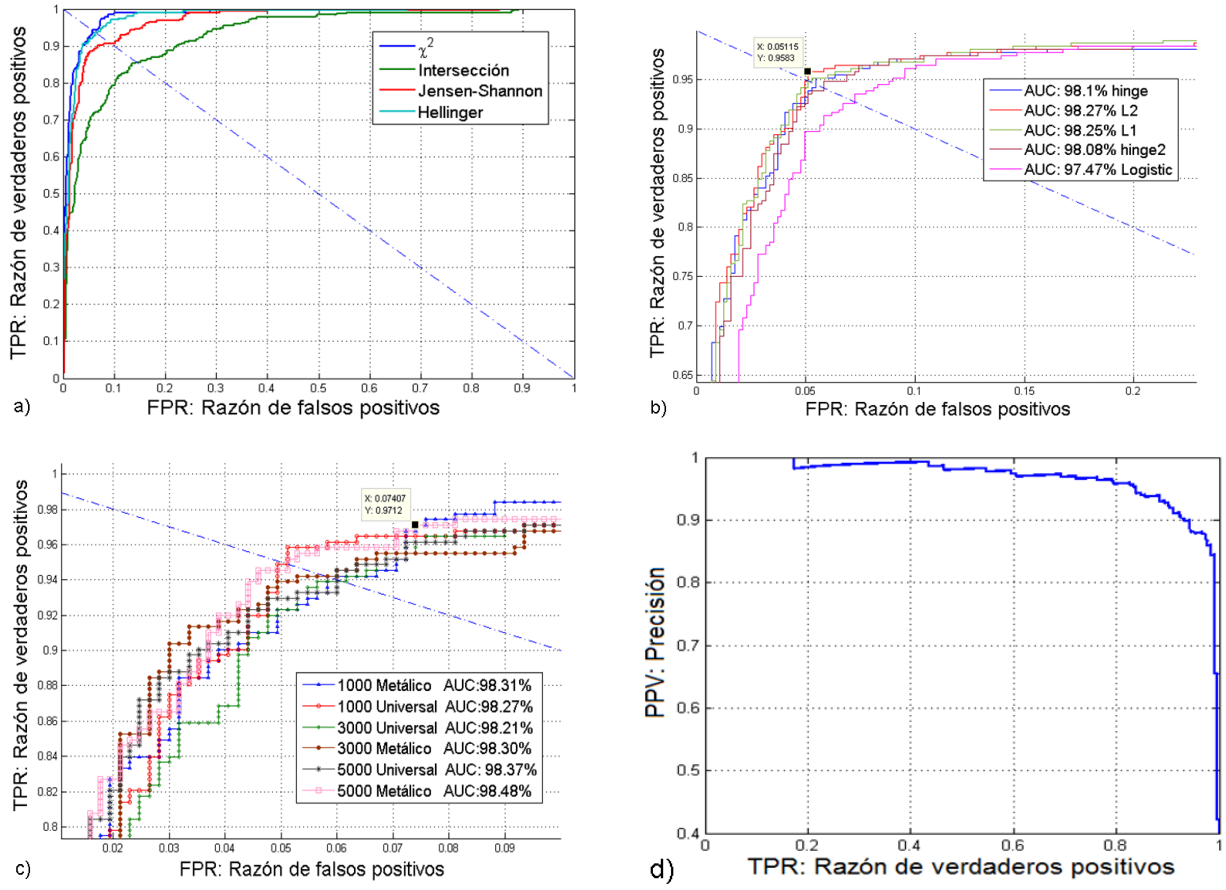

Figura 3. a) Curvas ROC de los kernels, b) Curvas ROC de las funciones de pérdida, c) Curvas ROC de los vocabularios y punto de operación, d) Curva *Precision-Recall*

Además se realizó un experimento para medir el comportamiento del algoritmo frente al reconocimiento de armas solapadas con objetos metálicos mediante el método de retención o (*holdout*). Se utilzaron todas las armas solapadas en el conjunto de prueba y las restantes en el conjunto de entrenamiento. Se obtuvo un resultado de TPR=88.83 % y FPR=12.7 %. Este resultado se puede tomar como la medida de la peor precisión que tiene el algoritmo frente al reconocimiento de armas con solapamiento.

Análisis y discusión

A manera de análisis de los resultados se presenta una comparación con los métodos de Başţan (BAŞTAN et al., 2011) (TPR 70 %, PPV 29 %) y Turcsany (TURCSANY et al., 2013) (TPR 99.07 %, FPR 4.31 %). También

se analizan los resultados del reciente trabajo de Baştan (BAŞTAN et al., 2013). Este trabajo (TPR=97.12 %, FPR=7.4 %) alcanza resultados cercanos al de Turscany, mejores resultados encontrados en la literatura. Sin embargo, el uso de diferentes bases de datos en cada trabajo hacen que la comparación no pueda hacerse estrictamente basada en los resultados. Las bases de datos difieren, principalmente en cuanto a cantidad de imágenes y puntos de vista de los objetos. Además debe tener en cuenta que las imágenes usadas por Turcsany son de equipos de una sola energía (escalas de grises) diferentes a las utilizadas en esta propuesta. Aunque la base de datos utilizada en este estudio no posee una cantidad de imágenes similar al trabajo de Turcsany, sí aparecen imágenes representativas de ambas clases en diversas situaciones de complejidad. En los estudios referenciados no aparece una descripción de este tipo.

Existe una diferencia conceptual implícita entre los trabajos referenciados en este análisis y es que Baştan en 2011 realiza una clasificación sobre imágenes completas (todo el equipaje) mientras el resto aplica el concepto de ventana deslizante, concentrándose en el objeto a detectar. Esto debe traer consigo una diferencia notable en los histogramas obtenidos. A opinión de los autores esta debe ser la razón principal de la diferencia en los resultados con el trabajo de Baştan en 2011. Las principales diferencias en los métodos se concentran en: la extracción de características, tanto en la detección de los puntos como en el descriptor, el tipo de vocabulario construido, el tipo de kernel junto con otros parámetros que se pueden modificar en el entrenamiento de la SVM y la experimentación con la base de datos.

En este trabajo se propone el uso de un vocabulario construido solo con características de objetos metálicos (vocabulario metálico). El mismo presentó mejores resultados que el construido con todo tipo de características (vocabulario universal). Esto se puede deber a que al filtrar los rasgos de zonas no metálicas, se elimina información no relevante en los histogramas, concentrándose en representar diferentes objetos metálicos. Por esta razón y por el número de palabras usadas se cree que el vocabulario propuesto es superior al de Baştan (BAŞTAN et al., 2011). Además se obtuvo que el kernel χ^2 es superior al Intersección, esto contradice lo mencionado por Baştan (BAŞTAN et al., 2011, 2013) donde se plantea que el Intersección es superior. Sin embargo en ninguna de las referencias consultadas aparece la gráfica que muestra la comparación de las curvas ROC de los kernels utilizados en busca del que tenga mejor comportamiento. Este experimento es de suma importancia dado la incidencia en el resultado general. Para otros tipos de imágenes también se ha reportado la superioridad de χ^2 (JIANG et al., 2007; ZHANG et al., 2007), lo que resulta coherente con nuestro resultado. Igualmente, en las referencias consultadas no aparecen los experimentos en busca de la mejor función de pérdida. Se pudo comprobar en este estudio cómo este parámetro influye en la calidad de los resultados. Adicionalmente se obtuvo una medida de la peor precisión del algoritmo en el reconocimiento de armas solapadas con objetos metálicos, análisis que no se contempla en demás estudios. Dado que es más probable encontrarse esta situación, como un intento para esconder armas, tiene gran importancia analizar la eficiencia del algoritmo para esta situación.

Por otro lado el trabajo de Baştan (BAŞTAN et al., 2013) utiliza la métrica AP (*Average Precision*) que es el área bajo la curva *Precision-Recall* para presentar sus resultados. Para poder realizar una comparación se calculó la curva *Precision-Recall* que aparece en la figura 3 d). Donde se obtuvo como resultado 96.48 % de AP. Este resultado supera al mejor resultado de 94.6 % que presenta Baştan en 2013 para el caso de clasificación con una única vista en armas cortas. El resultado de Baştan 2013 fue alcanzado con otro algoritmo de extracción de características, un vocabulario universal de 5000 palabras visuales, sin realizar validación cruzada y con el kernel de Intersección de histogramas.

Conclusiones

En este artículo se mostró el desarrollo de un algoritmo para el reconocimiento de armas cortas en imágenes de rayos X usando el método Saco de Palabras Visuales, alcanzando un resultado general con razón de verdaderos positivos de 97.12 % y razón de falsos positivos de 7.4 %. Se contribuyó en el tipo de vocabulario construido, el kernel y la función de pérdida utilizada. Para futuros trabajos investigar una alternativa para aumentar la eficiencia en el reconocimiento de armas solapadas con objetos metálicos. Los resultados alcanzados en esta investigación muestran que es posible implementar un sistema de visión por computadora que reconozca armas de fuego, que facilite el trabajo de los operadores y que el proceso de inspección sea más rápido y preciso en el reconocimiento de objetos peligrosos.

Referencias

BAŞTAN, M., BYEON, W., and BREUEL, T. M. (2013). Object recognition in multi-view dual energy x-ray images. In *British Machine Vision Conference BMVC*.

BAŞTAN, M., YOUSEFI, M. R., and BREUEL, T. M. (2011). Visual words on baggage x-ray images. In *Computer Analysis of Images and Patterns*, volume 6854 of *Lecture Notes in Computer Science*, pages 360–368. Springer Berlin Heidelberg.

BOSCH, A., ZISSERMAN, A., and MUNOZ, X. (2007). Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

CHATFIELD, K., LEMPITSKY, V., VEDALDI, A., and ZISSERMAN, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*.

CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J., and BRAY, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2.

GONZALEZ, R. and WOODS, R. (2002). *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.

JIANG, Y. G., NGO, C. W., and YANG, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501. ACM.

JONES, M. and VIOLA, P. (2001). Robust real-time object detection. In *Workshop on Statistical and Computational Theories of Vision*.

LOWE, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

PERRONNIN, F., DANCE, C., CSURKA, G., and B., M. (2006). Adapted vocabularies for generic visual categorization. In *Computer Vision–ECCV 2006*, pages 464–475. Springer.

TURCSANY, D., MOUTON, A., and BRECKON, T. P. (2013). Improving feature-based object recognition for x-ray baggage security screening using primed visualwords. In *Industrial Technology (ICIT), 2013 IEEE International Conference on*, pages 1140–1145. IEEE.

VAPNIK, V. N. (1998). *Statistical learning theory*, volume 1. Wiley New York.

VEDALDI, A. and FULKERSON, B. (2008). VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.

VEDALDI, A. and ZISSERMAN, A. (2011). Image classification practical. http://www.di.ens.fr/willow/events/cvml2011/ materials/practical-classification/. Accessed: 2014-05-10.

VEDALDI, A. and ZISSERMAN, A. (2012). Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492.

YANG, J., JIANG, Y. G., HAUPTMANN, A. G., and NGO, C. W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM.

ZHANG, J., MARSZAŁEK, M., LAZEBNIK, S., and SCHMID, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238.