Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

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

Detección de sombreros en imágenes de rostro con fondo uniforme

Detecting hats in face images with uniform background

Jesús Pérez-Martín[1*], **Yenisel Plasencia Calaña**[1]

[1]Centro de Aplicación de Tecnologías de Avanzada (CENATAV). {jmartin,yplasencia}@cenatav.co.cu

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

Resumen
La determinación de la calidad de una imagen de rostro es un paso importante para los métodos automáticos de reconocimiento de rostros, con el fin de que los algoritmos de reconocimiento de individuos a partir de imágenes de su rostro reciban como entrada imágenes de alto valor identificativo. Para lograr la detección de sombreros se propone la creación de un método de reconocimiento de patrones basado en un modelo de Bolsa de Palabras Visuales. Se realizaron pruebas con descriptores de rasgos SURF (del inglés Speeded Up Robust Features), SIFT (del inglés Scale Invariant Feature Transform), y los novedosos DSIFT (del inglés Dense SIFT) y PHOW (del inglés Pyramid Histogram of visual Words) que obtuvieron los mejores resultados. Además, proponemos ejecutar la clasificación mediante máquinas de vectores de soporte usando el kernel de intersección de histogramas. Este kernel, que hace relativamente poco tiempo se descubrió que cumple las propiedades necesarias para ser usado en el contexto de estos clasificadores, hace el papel de una similitud y es apropiado para tipos de datos como los calculados basados en histogramas. Los resultados experimentales muestran que se logra una alta eficacia en el problema abordado.

Palabras claves: modelo de Bolsa de Palabras Visuales, histogramas espaciales, SURF, SIFT, DSIFT, PHOW

Abstract
Determining the quality of a face image is an important step for automatic face recognition methods, for the purpose that the algorithms of recognition and identification of individuals from images of his face receive an image of high identifying value as input. To detect hats, a pattern recognition method based on bag of visual words model is proposed. Testing with SURF, SIFT and the novel DSIFT and PHOW descriptors, which obtained the best results. In addition, classification is performed by SVM using the histogram intersection kernel. Recently it was discovered that this kernel fulfill the necessary conditions to be used in the context of these classifiers. It plays the role of a similarity and it is appropriate for data types such as those calculated based on histograms. Experimental results show that a high accuracy in the current problem is achieved.

Keywords: *Bag of Visual Words model, spatial histograms, SURF, SIFT, DSIFT, PHOW*

Introducción

Si bien existen investigaciones sobre las variaciones en la pose de las personas, iluminación y degradación de la imagen, casi todos los enfoques existentes para el reconocimiento facial en condiciones de oclusión se centran en detectar el uso de gafas de sol y bufanda. La oclusión causada por sombreros no ha sido estudiada, a pesar de las ventajas que brindaría para el desempeño de los algoritmos de detección de rostro y el reconocimiento de la identidad, además de ser uno de los requisitos de calidad planteados por la ICAO para las imágenes de rostro (FERRARA et al., 2012).

Método

Para dar solución a este problema, en el presente trabajo se propone la creación de un método de reconocimiento de patrones, basado en un modelo de Bolsa de Palabras Visuales (BoVW), donde cada imagen va a estar representada por un conjunto de vectores en vez de por un solo vector de características como sucede en los enfoques clásicos.

Modelo de Bolsa de Palabras Visuales

BoVW es actualmente un método popular para el reconocimiento de objetos y escenas en visión por computadoras. A una imagen se le extraen los rasgos locales y pasa a ser considerada como una *bolsa de rasgos* (*bag of features*), es decir, ignorando las relaciones espaciales entre ellos. Como desventaja podemos mencionar que este no cuenta con un mecanismo eficiente y efectivo de codificación de la información espacial que existe para los rasgos. Un método basado en el BoVW clásico consiste en las siguientes etapas:

- **Extracción de rasgos**: Los rasgos locales y sus descriptores correspondientes se extraen de parches locales de la imagen. Los dos descriptores visuales más usados son SIFT (LOWE, 2004) y SURF (VEDALDI and FULKERSON, 2010). Algunos métodos los extraen en ciertos puntos de interés detectados y otros obtienen los rasgos locales densamente, en posiciones regulares de la imagen por ejemplo PHOW (VEDALDI and FULKERSON, 2010).
- **Generar un diccionario y mapear los rasgos a palabras visuales**: Un diccionario visual es un método que divide el espacio de descriptores visuales en varias regiones. Los rasgos de una región corresponden a la misma palabra visual. Entonces, una imagen se codifica como un histograma de la frecuencia de ocurrencia de cada palabra visual. Esto se hace asignando a cada vector de rasgos de la imagen su región más cercana, de manera que al terminar el proceso se tenga la cantidad de vectores asignados a cada región y se asigna esa cantidad a la componente correspondiente a esa palabra visual en el histograma.
- **Entrenar y probar** : Varios métodos de aprendizaje por computadora pueden aplicarse para la repre-

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

sentación de imágenes usada. SVM es frecuentemente usado como clasificador en modelos BoVW para el reconocimiento de objetos y escenas. Este fue el clasificador escogido para resolver el problema planteado, en conjunto con el kernel aditivo de intersección de histogramas debido a su utilidad y buen desempeño para representaciones basadas en histogramas.

Preprocesamiento

Como el problema se centra solo en imágenes de rostro de fondo uniforme y los sombreros siempre se encuentra en una misma región relativa a las personas, se decidió, con el fin de reducir el área de búsqueda, seleccionar de la imagen la región en la que debe estar el sombrero. Esto se hace convirtiendo la imagen a escala de grises y a partir de la detección del rostro de la persona, se realiza un escalado de manera que sus ojos queden a una distancia de 20 píxeles y finalmente se selecciona una región de la imagen que se extiende desde la mitad del rostro hacia arriba, con un ancho y alto no mayores de 100 píxeles. Si el rostro no fuese detectado, entonces la imagen es escalada a una altura de 200 píxeles. En la figura 1 se muestra el resultado de aplicar este proceso para una imagen de ejemplo.

La idea fundamental de el preprocesamiento propuesto es eliminar la influencia que puede tener la distancia a la que fue tomada la imagen y garantizar lo mejor posible que el sombrero sea segmentado completamente del resto de la imagen. La imagen también se lleva a escala de grises.

Figura 1. Preprocesamiento de la imagen, selección de la región de interés a partir de la detección del rostro y el escalado de la imagen

Extracción de características

Como parte de la investigación desarrollada en este trabajo, en la búsqueda de la mejor solución al problema de la detección de sombrero se probaron varios de los métodos que se mencionan en la literatura para la extracción de las características de las imágenes y que han mostrado buenos resultados, estos son SIFT, SURF, DSIFT

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

y su variante, PHOW.

Construcción del diccionario de Palabras Visuales

Una vez detectados los puntos claves y extraídas las características con su descriptor, BoVW propone la creación de un diccionario visual, un conjunto de palabras visuales, con el fin de describir posteriormente las imágenes mediante la detección de la ocurrencia de estas palabras en ellas. La calidad del diccionario visual tiene un impacto significativo sobre el éxito de los métodos basados en BoVW. Muchos métodos para la categorización de objetos y escenas emplean métodos de aprendizaje no supervisado (por ejemplo, el agrupamiento k-means) para obtener dicho diccionario visual, tomando como palabras visuales los centroides obtenidos para cada grupo en este proceso.

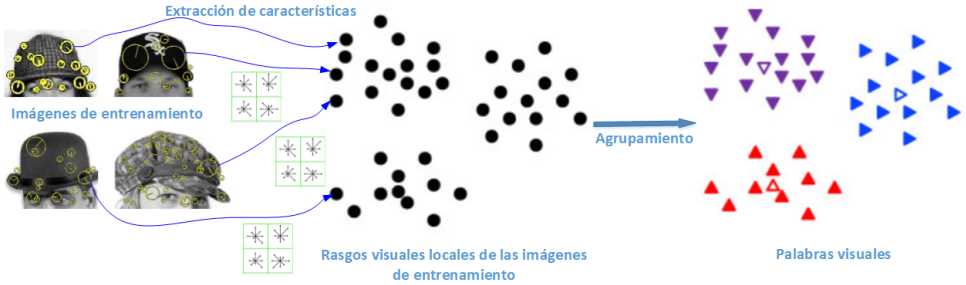

Figura 2. Proceso de creación del diccionario visual, extracción, representación y agrupamiento.

Se hicieron pruebas con agrupamiento k-means usando distancia Euclidiana y jerárquico aglomerativo con vinculación media (average linkage, en inglés) usando distancia χ^2, con el fin de determinar cuál se ajustaba mejor al problema. Un aspecto fundamental en el rendimiento de estos métodos es el tamaño del diccionario, puesto que esto influye en la capacidad representativa y discriminativa de las palabras visuales sobre las clases. A pesar de existir varios estudios sobre este tema, no existe ninguna regla para determinar a priori qué tamaño dará los mejores resultados. En artículos como (LAZEBNIK et al., 2010; BOSCH et al., 2007) se logran buenos resultados con una cantidad de palabras en el rango de entre las 100 y las 800.

Descriptores de las imágenes

Para describir las imágenes se utilizo un esquema de *Emparejamiento Piramidal Espacial* (*Spatial Pyramid Matching*), donde se plantea el cálculo de histogramas de frecuencias de las palabras visuales a distintas resoluciones de la imagen. En algunos artículos como (LARA and Jr., 2011; HADJIDEMETRIOU et al., 2004) las diferentes resoluciones se determinan mediante repetidos submuestreos de la imagen y computan un histograma global de los valores de los píxeles para ese nivel, se varía la resolución a la que los rasgos son

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

calculados (valores de los píxeles), pero la resolución del histograma (escala de intensidad) se mantiene fija. En artículos como (GRAUMAN and DARRELL, 2005; LAZEBNIK et al., 2010) se plantea un enfoque opuesto, fijar la resolución a la que se determinan los rasgos, pero variar la resolución espacial en la que son agregados. En este trabajo se propone un Emparejamiento Piramidal Espacial variando ambas resoluciones.

Para entender claramente esta estrategia y en qué criterios se basan sus buenos resultados, primeramente se expondrá la formulación original del kernel de *intersección de histogramas* y de *emparejado piramidal* (GRAUMAN and DARRELL, 2005) y luego se introduce su aplicación en la representación de la imagen en la solución propuesta.

Kernel de Intersección de Histogramas

Sea $\mathbf{x} = (x_1, x_2, ..., x_d) \in \mathbb{R}_+^d$ un histograma de valores reales no negativos de d intervalos. $\mathbf{x}$ pudiera representar una imagen (como en la formulación clásica del modelo de bolsa de palabras visuales) o un parche de una imagen (como los descriptores SIFT). El *kernel de intersección de histogramas* K_{HI} se define como (WU et al., 2011):

$$K_{HI}(\mathbf{x_1}, \mathbf{x_2}) = \sum_{j=1}^{d} \text{mín}(x_{1,j}, x_{2,j}) \tag{1}$$

Emparejamiento Piramidal Espacial

Sea X y Y dos conjuntos de vectores en un espacio de características d-dimensional. Grauman y Darrell (GRAUMAN and DARRELL, 2005) proponen el *emparejado piramidal* para encontrar una correspondencia aproximada entre estos dos conjuntos. Informalmente, la idea trabaja mediante la distribución del espacio de características en una secuencia de rejillas que van aumentando su número de celdas (niveles de la pirámide) y el cálculo de la suma ponderada de la cantidad de correspondencias detectadas en cada nivel. Se dice que dos puntos corresponden en un mismo nivel, si caen ubicados dentro de la misma celda de la rejilla. Las correspondencias que se encuentran en un nivel más bajo, se ponderan con un mayor valor que las que se encuentran en los primeros niveles. Específicamente, se construye una secuencia de rejillas de resolución $0, ..., L$, tal que el número de subregiones (celdas) en el nivel l es de 2^l por cada dimensión, para un total de $D = 2^{dl}$ subregiones. Sean H_X^l y H_Y^l los histogramas de X y Y en el nivel l, entonces $H_X^l(i)$ y $H_Y^l(i)$ representan el número de puntos de X y Y que están dentro de la i-ésima celda en el nivel l de resolución. Entonces el número aproximado de emparejamientos entre X y Y en el nivel l se determina mediante la *intersección de*

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

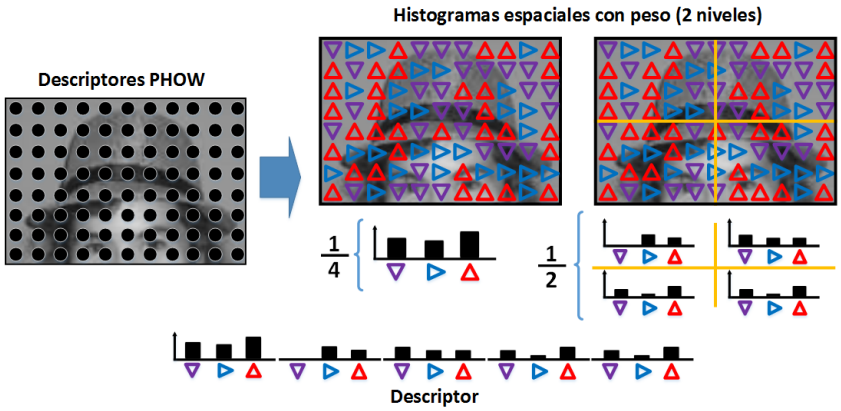

Figura 3. Creación del los descriptores para las imágenes: se extraen los descriptores, luego a cada uno se le asigna la palabra más semejante, posteriormente se calculan los histogramas espaciales (en el ejemplo se usa un nivel de profundidad). El descriptor final se determina a partir de la concatenación de todos los histogramas calculados, asignándole un peso según el nivel

histogramas a partir de la fórmula 1 de la siguiente manera:

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} \text{mín}(H_X^l(i), H_Y^l(i)) \tag{2}$$

abreviadamente I^l

El número de emparejamientos encontrados en el nivel l incluye a los encontrados en el nivel $l+1$. Por tanto la cantidad de nuevos emparejamientos está dada por $I^l - I^{l+1}$ para $l = 0, ..., L-1$. El peso asociado con el nivel l será $\frac{1}{2^{L-l}}$, el cual es inversamente proporcional al tamaño de las subregiones del nivel. Con lo que se busca penalizar las correspondencias encontradas en los niveles de las celdas más grandes, ya que en estos se incrementa el número de características disimilares. Se define un *kernel de emparejado piramidal* como:

$$\begin{aligned} K^l(X,Y) &= T^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^{l+1}) \\ &= \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}(I^l) \end{aligned} \tag{3}$$

Ambos kernels son funciones simétricas y definidas positivas, cumplen con los requisitos del Teorema de Mercer (GRAUMAN and DARRELL, 2005) y por tanto son kernels válidos para usar con un modelo SVM.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

Finalmente, para determinar el descriptor de la imagen, dado un diccionario de M palabras visuales, como el kernel de emparejamiento piramidal (ecuación 3) es simplemente una suma ponderada de intersección de histogramas y ya que $c\,\text{mín}(a,b) = \text{mín}(ca,cb)$ para números positivos, se puede describir la imagen como un largo vector formado por la concatenación de los histogramas normalizados de todas las resoluciones ponderados apropiadamente y calcular K^L como la intersección de estos largos histogramas (LAZEBNIK et al., 2010). El vector resultante tendrá dimensión $M\sum_{l=0}^{L} 4^l$. En la figura 3 se muestra un ejemplo de este proceso.

Clasificación

El resultado de este proceso es la decisión sobre la clase a la que pertenece la imagen. Cada imagen es reconocida como perteneciente a uno de los siguientes tipos: Personas con sombrero, Personas sin sombrero. Existen varias técnicas de clasificación que han sido probadas para el modelo BoVW y han mostrado buenos resultados. En este trabajo primeramente se desarrolló una idea basada en la aplicación de una red Bayesiana de tres niveles. Posteriormente se decidió cambiar esta idea por la creación de un modelo SVM dado que mostró mejores tasas de clasificación para el problema de dos clases que se plantea.

Experimentos

Para crear los diccionarios visuales se seleccionaron 30 imágenes aleatorias del conjunto de entrenamiento, 20 de la clase *persona con sombrero* y 10 de la clase *persona sin sombrero*. A estas se le extrajeron los rasgos usando SURF, SIFT, DSIFT y PHOW, tomando por cada tipo de rasgo hasta un total de 100000 descriptores aleatorios para construir el diccionario visual mediante el algoritmo de agrupamiento *k-means*, fijando como cantidad de grupos (palabras visuales) a generar, $M = 100, 200, ..., 600$. Para un total de $4\times6 = 24$ diccionarios.

A partir de los 24 diccionarios generados, es necesario determinar cuál se ajusta mejor al problema, seleccionando así, el método de extracción de características y la cantidad de palabras que usará la solución finalmente. Luego se realizó otro experimento para determinar el valor de la constante C para el modelo SVM a generar y la cantidad de niveles a tener en cuenta en el análisis espacial de modo que se minimice el por ciento de error. Para conseguir esto se realizaron las siguientes etapas:

- Se seleccionaron 170 imágenes con un balance entre la cantidad por clase.
- De las imágenes se extraen 4 descriptores por cada uno de los 24 diccionarios, los cuales representan los niveles de profundidad en el análisis espacial, $l = 0, 1, 2, 3$. Para un total de $4 \times 24 = 96$ representaciones de las 170 imágenes.
- Para cada uno de ellos se realizaron 10 divisiones aleatorias sucesivas al 50 % usando una mitad para entrenar 4 modelos SVM con kernel de intersección de histogramas y valores distintos del parámetro

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

$C = 0.1, 1, 10, 100$ y la otra para probar, siempre garantizando que las dos mitades queden balanceadas en cuanto a la cantidad de imágenes por clases.

En el gráfico 4 se muestra las combinaciones que mejor precisión alcanzaron para cada uno de los métodos de extracción de características.

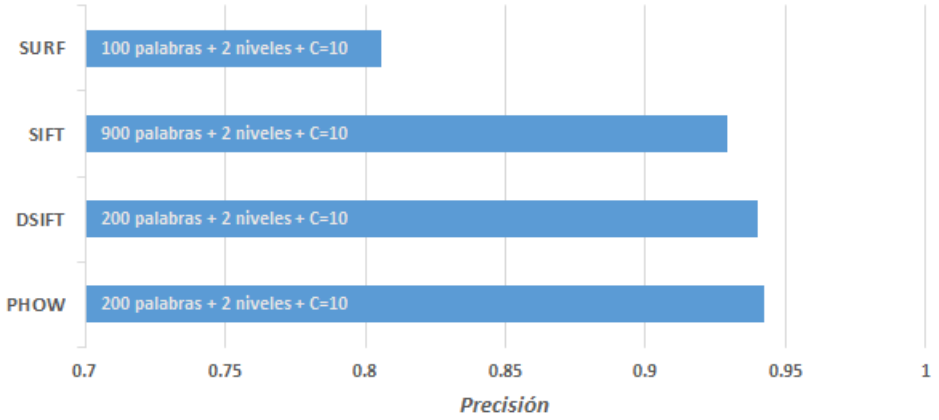

Figura 4. Mejores resultados de los experimentos para los métodos de extracción de características SURF, SIFT, DSIFT y PHOW.

Un resumen de los resultados obtenidos para los diccionarios generados mediante el método PHOW hasta nivel 2 de profundidad se representa en el gráfico 5, donde se observa hasta aproximadamente un 6 % de error como promedio en las 10 divisiones aleatoria en algunos casos, así como para diccionarios pequeños es mejor seguir un enfoque espacial mientras para los más grandes analizar solo la imagen completa es lo más conveniente.

A partir de estos resultados, se procedió a la clasificación de las restantes 100 imágenes usando solo PHOW, para determinar el método de extracción de características, la cantidad de palabras y el nivel de profundidad para el análisis espacial de mejor precisión. La tabla 1 muestra los resultados obtenidos para estos experimentos.

Discusión

Analizando los valores mostrados en la tabla 1 para PHOW, se nota que procesar la imagen de manera global (nivel 0), así como descender demasiado (hasta nivel 2), produce resultados poco estables, obteniéndose hasta una precisión por debajo del 90 %. Mientras que para el nivel 1, se registran valores de precisión más estables para todos los tamaños del vocabulario visual, superando en todos los casos el 91 %. A su vez, se tiene que el

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

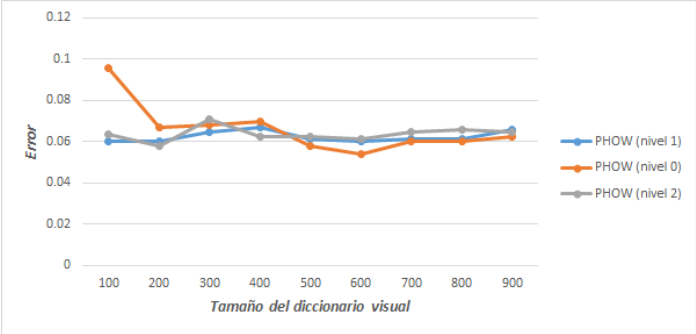

Figura 5. Menor error obtenido de los cuatro modelos SVM (con distintos valores del parámetro C) probados para cada uno de los tamaños del diccionario generados mediante la variante PHOW

Tabla 1. Resultados de clasificar 100 imágenes usando PHOW con diccionarios desde 100 hasta 600 palabras, usando $l = 0, 1, 2$ niveles de profundidad en el análisis espacial y con parámetro $C = 10$ para el SVM.

Nivel	**Cantidad de palabras**					
	100	**200**	**300**	**400**	**500**	**600**
0	91 %	94 %	92 %	89 %	91 %	91 %
1	91 %	94 %	93 %	90 %	94 %	92 %
2	88 %	93 %	92 %	92 %	93 %	91 %

tamaño del vocabulario con mejores índices de precisión es $M = 200$, alcanzando hasta 93 % y mostrando los mejores valores para los primeros niveles, factor decisivo en la eficiencia ya que estos son los que operan con vectores de menor dimensión y por tanto generan un menor número de operaciones.

Si a esto se añade que para la variante PHOW, el gráfico 5 muestra que el nivel 1 fue el más estable, reportando el segundo valor mínimo de error general con un tamaño del vocabulario visual de $M = 200$; se justifica proponer como solución final un método de reconocimiento de patrones basado en BoVW con la siguiente configuración:

1. PHOW como método de extracción de rasgos de las imágenes.
2. Diccionario visual de tamaño $M = 200$.
3. Describir las imágenes mediante vectores de 1000 componentes, formados a partir de la concatenación de los histogramas locales obtenidos de la división de la imagen en 4 regiones a lo sumo (descender en el análisis espacial hasta un nivel de profundidad $l = 1$), para dar una respuesta lo más rápido posible sin perder demasiado en precisión.
4. Finalmente clasificar las nuevas imágenes a partir de un modelo SVM con kernel de intersección de histogramas con parámetro de holgura $C = 10$ y previamente entrenado con el conjunto de 170 imágenes.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 261-270

Conclusiones

Usando el enfoque de BoVW con clasificación basada en SVM se desarrolló un nuevo método eficaz para detectar la presencia de sombreros en imágenes de rostro con fondo uniforme, brindando una solución completa para este problema que ya ha sido implementada en lenguaje $C++$ y agregada a uno de los proyecto aplicados del CENATAV que busca determinar de la calidad de las imágenes de rostro y por tanto el valor identificativo que poseen.

Referencias

BOSCH, A., ZISSERMAN, A., and MUNOZ, X. (2007). Image classification using random forests and ferns. In *IEEE International Conference on Computer Vision*.

FERRARA, M., FRANCO, A., MAIO, D., and MALTONI, D. (2012). Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*.

GRAUMAN, K. and DARRELL, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *IN ICCV*, pages 1458–1465.

HADJIDEMETRIOU, E., GROSSBERG, M., and NAYAR, S. (2004). Multiresolution Histograms and Their Use for Recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26:831–847.

LARA, A. C. and Jr., R. H. (2011). Combining features to a class-specific model in an instance detection framework. In Lewiner, T. and da Silva Torres, R., editors, *SIBGRAPI*, pages 165–172. IEEE Computer Society.

LAZEBNIK, S., SCHMID, C., and PONCE, J. (2010). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178. IEEE Computer Society.

LOWE, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, pages 91–110.

VEDALDI, A. and FULKERSON, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472. ACM.

WU, J., TAN, W.-C., and REHG, J. M. (2011). Efficient and effective visual codebook generation using additive kernels. *J. Mach. Learn. Res.*, pages 3097–3118.