# Object Detection with Vocabularies of Space-time Descriptors

Y. Hernandez-Heredia*[1], J.M.[a] González-Linares[3], N. Guil[3], J. Ortiz[2], R. Hernandez[1], J.R. Cózar[3]

[1] Centro de Geoinformática y Señales Digitales
Universidad de las Ciencias Informáticas
Cuba, Habana
*yhernandezh@uci.cu
[2] Vicerrectoria de Tecnología
Universidad de las Ciencias Informáticas
Cuba, Habana
[3] Departamento de Arquitectura de Computadores
Universidad de Málaga, E.T.S.I. Informática
España, Málaga

**ABSTRACT**
This paper presents a novel framework for objects detection in security and broadcast videos. Our method assumes that object classes are unknown in advance and exploit the temporal-space properties of the videos for the creation of a vocabulary that describes these classes. Local space-time features have recently became a popular video representation for action recognition and object detection. Several methods for feature localization and description have been proposed in the literature and promising recognition results were demonstrated for a number of action classes.

In this work we propose the use of different kinds of descriptors for the creation of vocabularies for different detection object task. For a better description of the videos we carry out a background model, tryring to clean up and follow the areas where there are objects. The points of interest in the videos to characterize the objects are calculated with a temporary variant of the famous Harris corner detector. With the descriptors obtained from the points of interest, a vocabulary is realized usingthe kinds of videos we want to train. Then we obtained the frequency histograms between the videos for training and the vocabulary so, with a binary classifier obtain the trained classes and following the same procedure without the vocabulary realized the detection and monitoring of the objects.

The new method presented is also compared with a state of the art method, obtaining better results in both accuracy and false object rejection.

Keywords: object detection, video segmentation, vocabulary, binary classifier.

**RESUMEN**
Este artículo presenta un método novedoso para la detección de objetos en videos de seguridad y de transmisión de televisión. Nuestro método supone que las clases de objetos son desconocidas por adelantado y explota las propiedades temporales y espaciales de los videos para la creación de un vocabulario que describe estas clases. Las características locales del espacio y el tiempo se han convertido recientemente en una representación popular de los vídeos para el reconocimiento de acciones y la detección objetos. En estudios recientes se han propuesto varios métodos para la localización y descripción de características de videos y han demostrado resultados prometedores de reconocimiento para clases de acción de personas y objetos.

En este trabajo proponemos el uso de diferentes tipos de descriptores para la creación de vocabularios para tareas de detección de objetos diferentes. Para una mejor descripción de los videos generamos el modelo del fondo para tratar de limpiar y seguir las zonas donde están los objetos. Los puntos de interés de los videos para caracterizar a los objetos se calculan con una variante temporal del famoso detector de esquinas Harris. Con los descriptores obtenidos de los puntos de interés se realiza un vocabulario con las clases de videos que se quieran entrenar. Luego se obtienen los histogramas de frecuencia entre los videos de entrenamiento y el vocabulario para con un clasificador binario obtener las clases entrenadas y siguiendo el mismo procedimiento sin el vocabulario realizar la detección y seguimiento de los objetos.

El nuevo método presentado también se compara con propuestas actuales para situaciones similares, obteniendo mejores resultados en la precisión y el rechazo de objetos falsos.

## 1. Introduction

There are several techniques in the world literature for objects detection in videos and images [1, 2], however, they usually do not have good results when used in real applications of analysis of video such as video surveillance or monitoring of television signals. Most of the approaches [3, 4] segment the frames taken by the cameras as a first step, by identifying the background[1] of the scene and then, identifying, the foreground[2] compound by moving objects [5]. Afterwards, the techniques include tracking algorithms to analyze the evolution of objects. The objects detected using these techniques are depicted as blobs[3] that identify the area of the image occupied by the object.

The recognition of these objects requires the use of advanced techniques that combine three essential elements to optimize the expected results in environments that are not controlled; taking into account changes in perspective, lighting and colors, as well as errors in the image that might appear, introduced by the encoding of videos:

1. The correct selection of the features that are used to represent the object.
2. The compact representation of these features through descriptors.
3. The proper construction of a model of the object allowing to assimilate, conveniently, changes of form in the same lighting changes, rotations, scaled and perspective transformations, as well as to make it robust to errors and artifacts that appear to encode videos.

Using the temporary information of scenes (description of the scene actions, motion, making changes, camera shots), coupled with spatial information (relationship between the elements of the scene, next to what), allows us to improve the descriptors with semantic labeling for the object information.

The work presented in this paper circumvents previous problems by pre-processing the video and by a correct selection of descriptors for different tasks. Thus, frames containing noisy objects are safely rejected without compromising the technique accuracy. Then, sequences with similar objects in the space are trained whit the correct words in a cluster to obtain the best possible classification. As a result of this identification, the temporal location of the objects in the video are detected.

The rest of the paper is organized as follows: Section 2 introduces a new technique for object detection by training the vocabulary of space-time descriptors. In section 3 the technique is tested with a set of videos and compared against another approaches. Finally, in section 4 conclusions and future works are presented.

## 2. A new technique for object detection by training vocabulary of space-time descriptors

The capture of the characteristics of the space and time describes the forms and the movement in a video, also provides an independent representation of events respect to space-time changes and changes in scale, differences of backgrounds and multiple movements in a scene (see figure 1). These features tend to be drawn directly from the video and thus avoid possible errors of a method of pre-processing as the segmentation of motion and follow-up.

Figure 1. Images with a detected type car in different perspectives. (a) Controlled environment. (b) Front car and attenuation of shadows in darkness. (c) Several cars in different sizes and positions.

[1] Static content where is the additional information in the foreground.
[2] It contains the largest amount of information that identifies the video sequence.
[3] Spots on detected objects for tracking in video surveillance systems

Representation, detecting and learning are the main problems to be addressed in the design of a visual system for the recognition categories of objects. The challenge of detecting is the definition of metrics and algorithms which are suitable for matching models to images in the presence of the occlusion.

Some authors [6, 3] focus their works in the differences of the background and standardize the training examples moreover, the recognition often proceeds by an exhaustive search for the image position. Probabilistic approaches [7] with random models, where several pieces are combined, produced the principles and efficient methods of detection. The author [8] proposes an algorithm with a high likelihood of learning unsupervised for different categories of objects which is an example of the raised previously.

Figure 2 shows the procedure to be followed with the videos of entry into the framework proposed for objects detection.

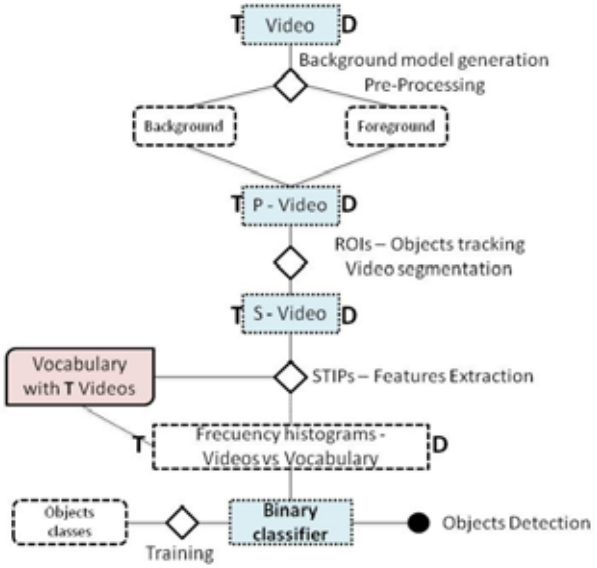


Figure 2. Object detection procedure. T – Training Videos. (D) Videos whit objects to be to detect.

Two groups of classes of videos, training sequences and detection are used for the model. Apply a pre-processing for separating the background of the image where the objects to be classify are, and thereby gain accuracy and time in the subsequent main processing of extraction of STIPs[4] and descriptors. Once we have the arrays of descriptors of each video, they are used for training to create a vocabulary with a Kmean, and then the histograms of frequency of the videos (training and detection) with respect to vocabulary, there are two groups of data needing the binary classifier to train and detect possible classes of objects that you have and want.

*2.1 Descriptors and STIPs extraction*

After obtaining the areas of interest of the videos analyzed during the pre-processing where the objects are to follow, the goal is to get the points of interest which they characterize and define them as objects. Currently, there are several algorithms to detect points of interest, some variants based on techniques to detect points in images, such as Harris [9] or Hessian [10] and others that use the space and time directly to identify points of interest in video, as the detector Cuboid sequences [11].

To model a space-time image sequence $f$, builds its linear scale-space representation by the convolution of $f$ with an anisotropic[5] Gaussian Kernel and different variances spatial $\sigma_l^2$ and temporary $\tau_l^2$.

$$L(\cdot;\sigma_l^2,\tau_l^2) = g(\cdot;\sigma_l^2,\tau_l^2) * f(\cdot) \quad (1)$$

Space - time Gaussian weight function is then a matrix of second-order space temporary 3 x 3, composed by the average of the first derivatives.

$$\mu = g(\cdot;\sigma_l^2,\tau_l^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \quad (2)$$

Finally, to detect interest points, regions are searched in the function $f$ that has significant values $\lambda_1\lambda_2\lambda_3$ of $\mu$.

$$H = det(\mu) - k * trace^3(\mu) \quad (3)$$

[4] Interest Points of a video

[5] Mathematical operator which transforms two functions f and g in a third function that in a sense represents the magnitude which overlap $(f)$ and a relocated and inverted $g$ version

$$H = \lambda_1\lambda_2\lambda_3 - K(\lambda_1 + \lambda_2 + \lambda_3)^3 \quad (4)$$

Once we have the points of interest that best describe the videos, it´s necessary to extract the descriptors with the pixels from each image of the form $(x, y, t)$ that better results can cast according to the types of objects in motion we are seeking, for the results of this article according to the database used, the better descriptor was the MoSIFT[6], optical flow of the SIFT variant, also testswere realized using descriptors such as eSURF, histogram oriented of gradients (HOG) [12] and histogram of optical flow (HOF), as well as a variant of a vector with HOG/HOF joined [6].

*2.2 Vocabulary and histograms of frequency*

The aims of this main step are to create a vocabulary with the precise words to be compared with video sequences, and how to produce the histogram that best describes. The vocabulary known as a Bag of Words (BoW) [13, 14], is a technique used in various fields such as processing natural language information retrieval and analysis of patterns [15], consisting of the representation of a document using a set not ordered with the frequencies of occurrence of the words in a dictionary contained in such document.

The characteristics are used for the creation of the vocabulary (may be all or a number that has the best results according to tests) taken from the videos in the case of application. These features create a cluster of training comprising similar descriptors to obtain a named vocabulary. Then a Kmean[7] is used in the creation of the vocabulary with a number of words or counter in the cluster, equal to the following formula, which during all the tests neither gave the best results, without overloading, nor leaves below each word on the number of descriptors that conform, as well as 6 for more effective cluster Kmean executions:

$$BOWKMeansT = Cant_{Descript} * 0.04 \quad (5)$$

$$Cant_{Descript} = 0.3 * Total_{Descript} \quad (6)$$

Some extensions can be applied to improve outcomes for example:

- Delete all too common visual words (Stop Word Removal)
- Most informative visual words selection based on the frequency of occurrence in all documents, or the correlation between a Word and a class of documents (using statistics,$x^2$ gain information or mutual information).
- Use spatial information taking into account the position of the descriptor in the image (geometric restrictions) use visual bigrams to indicate the spatial proximity of two different words (using histograms of co-occurrence) [15].

The vocabulary trained with the cluster that is carried out and saved, creates histograms of frequency from each of the videos used in the model (training and detection). To achieve these histograms, which are not more than vectors with dimension equal to the number of words that indicate in each position that resembles the cluster, this video applies a hierarchical matching between key words and vocabulary. The Radius-Match matching which is the best one, is used for this test for each descriptor consultation that has less distance given a threshold, this step majkes sure to eliminate the vector, occurrences far removed from the cluster with the Descriptor-Matcher BruteForce.

*2.3 Detection of objects in a video sequence*

When you have two sets of videos (training and detection) frequency histograms, are passed to train a supervised classifier for having labeled classes and make the best possible detection.

For this step, trees binary search could be used as classifier, with very good results in [16] or Bayesians points machines [17], however, the data in

[6] http://lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm

[7] Analysis method of cluster which aims the partition of n observations on k groups in which each observation belongs to the Group closest to the average of all the cluster.

the vector are very regular (32 bit float), and is much easier to implement a support vector machine (SVM) widely used in vision by a computer by its various application forms and variants of implementation [18].

Support vector machines are a set of supervised learning algorithms developed by Vladimir Vapnik and his team at AT & T Labs [19].

The use of tools, libraries and applications, is currently very common with support for various programming languages or own interfaces from files with data training and regression [20]. These libraries are very easy to use and give users many options. For tests with the framework, we used the LibSVM of [21] with multi-class mode one against all for training and a pre-calculated Kernel:

$$K(H_i, H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in}-h_{jn})^2}{(h_{in}+h_{jn})}\right) \quad (7)$$

Where $H_i$ and $H_j$ are the frequency histograms and $V$ the vocabulary size.

### 3. Experimental Results

To test our method, we use databases of videos of actions, because our technology is designed for scenes and real videos, objects in images are not our intention, as it was explained earlier the descriptors that build vocabulary depend on the temporal information, and objects that are moving in real time are therefore needed. Our first test was with the database KTH[8], see table 1 and figure 3, widely used for testing initial methods because is very simple, contains classes of actions of people in controlled environments and small resolutions.

To analyze the relationship between descriptors and the amount of vocabulary words, the following graphs show us how accuracy varies by wisdom of the classes in the KTH database, as well as the confusion between them by similarity of descriptors which are very similar actions.

| | HOG/ HOF | HOG 3D | Mo-SIFT | Cubo-id | E-SURF |
|---|---|---|---|---|---|
| Harris 3D | 91.8 % | 89.0 % | - | - | - |
| Cubo-id | 88.7 % | 90.0 % | - | 89.1 % | - |
| Hessian | 88.7 % | 88.3 % | - | - | 81.4 % |
| Mo-SIFT | 89.5 % | 84.28 % | 91.83 % | - | - |
| VHarris[9] | 92.13 % | - | 92.02 % | - | - |

Table 1. Average precision of several methods using combinations of detectors/descriptors in the KTH database.

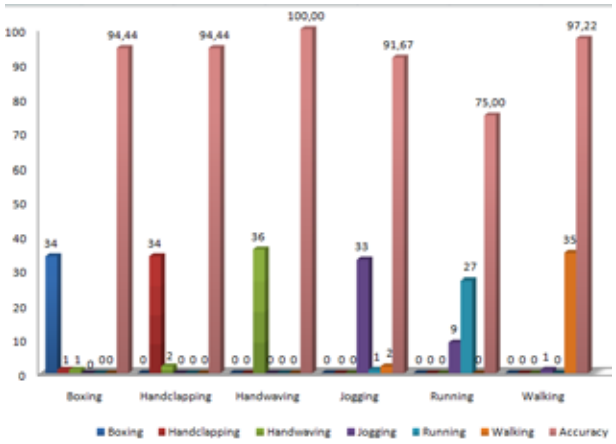


Figure 3. Results by classes of KTH Database. Boxing-Handclapping-Handwaving-Jogging-Walking.

| DET | DES | DES-A | BW | KM-Exec | ACC |
|---|---|---|---|---|---|
| VHarris | HOG / HOF | 30% | 4% | 6 | 92.1% |
| VHarris | HOG / HOF | 60% | 1% | 3 | 89.5% |
| VHarris | HOG / HOF | 100% | 2% | 2 | 75.6% |
| Mo-SIFT | Mo-SIFT | 60% | 2% | 8 | 83.7% |
| Mo-SIFT | Mo-SIFT | 30% | 4% | 4 | 89.2% |

Table 2. Different test changing the relationship between descriptors and clusters of the vocabulary.

[8] http://www.nada.kth.se/cvap/actions/

[9] Our Approach. Vocabulary with Harris3D Detector and the HOG/HOF - MoSIFT descriptors.

In the previously table, DET is the detector, DES, the descriptor, DES-A, the descriptors amount of the total, BoW, the descriptors amount for words in the vocabularies. KM-Exec, the total executions of the cluster and the ACC is the accuracy obtained.

Afterwards, the model was tested with more complex database (Hollywood Dataset[10]) videos with resolutions and real environments. This database contains 10 different classes with different scenes from movies with a lot of movement and diversity of background, the following table shows the results compared with the previous methods. See table number three.

The model also was tested with conventional cars and animals objects in normal videos that were perfectly detected by over 80% accuracy.

| | HOG/ HOF | HOG 3D | Mo-SIFT | Cubo-id | E-SURF |
|---|---|---|---|---|---|
| Harris 3D | 45.2% | 43.7% | - | - | - |
| Cubo-id | 46.2% | 45.7% | | 45.0% | - |
| Hessian | 46.0% | 41.3% | - | - | 38.2% |
| VHarris | 47.9% | - | 49.2% | - | - |

Table 3. Average precision of several methods using combinations of detectors/descriptors in the Hollywood database.

## 4. Conclusions

In this paper, we have proposed a method for learning the spatial and temporal structure of a visual object category in order to recognize new objects in this category, localize them in cluttered real-world scenes, and automatically obtain the segments from background. We have provided efficient algorithms for each of those steps and the resulting performance of recognition in several sets of data have been evaluated. Our results show that the method works well in different objects categories at different scales and achieves good performance of segmentation and detection of objects in difficult real scenes.

An important contribution of our work is the integration of an important segmentation of videos with the appropriate selection of descriptors that can characterize the best possible objects classes, as well as the creation of a vocabulary with the exact words for a proper matching with the training videos and videos test. Thus, the initial phase of recognition not only initializes the process of segmentation with the location of a possible object, but also gives an estimate of local measurements and its influence on the hypothesis of the object.

This mechanism constitutes a fundamental novelty in the detection of objects in real videos and improves results in the more precise acceptance decisions of conventional criteria based on the spatial characteristics of the images. This approach is flexible enough to be able to combine information from the descriptors according to the type of videos with the number of vocabulary words and type of matching to be use. The run time of the resulting approach mainly depends on three factors: model complexity (variation of the objects respect to the background), the size of the analyzed video (dimensions), and the selected search scale range.

Possible extensions include the integration and combination of several detectors of discrimination multi-categories and fusion of descriptors that are best suited to change perspective, lighting and colors using the type of material analyzed. Finally, you could also incorporate tests with other binary classifiers for the training and detection, as well as a cascade support vector machine.

[10] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

### *References*

[1] Yingzi, D. Unsupervised approach to color video thresholding. s.l. : Optical Engineering, 2004.

[2] Alfredo, M. Vision AIBO, ITAM. 2008.

[3] Cipolla, R., et al. Semantic texton forests forimage categorization and segmentation. s.l. : Computer Vision and Pattern Recognition, CVPR08. IEEE Conference, 2008.

[4] Laptev, I. histograms, Improving object detection with boosted. INRIA Rennes, France : Image and Vision Computing, 2009.

[5] Fergus, R., et al. Object class recognition by unsupervised scale-invariant learning. Oxford, UK : Computer Vision and Pattern Recognition, CVPR03 IEEE Computer Society Conference, 2003.

[6] Laptev, I., et al. Learning realistic human actions from movies. Anchorage, Alaska, USA : Computer Vision and Pattern Recognition, CVPR08., 2008.

[7] Burl, M.C., et al. probabilistic approach to object recognition using local photometry and global geometry. s.l. : Computer Vision, ECCV98, 1998.

[8] Weber, M., et al. Unsupervised Learning of Models for Recognition. Dublin, Ireland : Computer Vision - ECCV 2000, 2000. 978-3-540-67685-0.

[9] Laptev, I. and Lindeberg, T. On Space-time interest points. Springer : International Journal of Computer Vision, Kluwer Academic Publishers, 2005. 0920-5691.

[10] Willems, G. and Tuytelaars, T. An efficient dense and scale-invariant spatio-temporal interest point detector. s.l. : Lecture Notes in Computer Science, 2008. 978-3-540-88685-3.

[11] Dollar, P., et al. Behavior recognition via sparse spatio-temporal features. USA : Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2nd Joint IEEE International Workshop, 2005.

[12] Triggs, B. and Dalal, N. Histograms of oriented gradients for human detection. San Diego, CA, USA : Computer Vision and Pattern Recognition, CVPR05, 2005.

[13] Lazebnik, S., et al. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. New York, NY, USA : Computer Vision and Pattern Recognition, CVPR06, 2006.

[14] Gool, L.V. Bag of visual words model: recognizing object categories. England : Oxford University, 2007.

[15] Wallach, H.M. Topic Modeling: Beyond Bag-of-Words. New York USA : ICML '06 Proceedings of the 23rd international conference on Machine learning, 2006. 1-59593-383-2.

[16] Lempitsky, V. and Gall, J. Class-Specific Hough Forests for Object Detection. Miami, FL, USA : Computer Vision and Pattern Recognition, CVPR09, 2009.

[17] Herbrich, R. Bayes Point Machines. Department of Engineering Mathematics, Bristol University, United Kingdom : The Journal of Machine Learning Research, 2001.

[18] Rüping, S. mySVM - a support vector machine. 2010.

[19] Dragonfly Interactive. Nec Laboratories, INC America. 2008.

[20] Thorsten, J. Support Vector Machine for Complex Outputs. 2008.

[21] Chang, C. and Lin, C. LIBSVM -- A Library for Support Vector Machines. s.l. : ACM Transactions on Intelligent Systems and Technology (TIST), 2011.