REGIONS OF INTEREST FOR ACCURATE OBJECT DETECTION

P. Kapsalas, K. Rapantzikos, A. Sofou, Y. Avrithis

Image Video and Multimedia Systems Laboratory,
Department of Electrical & Computer Engineering,
Athens 15780, Greece
{pkaps, rap, natasa, iavr}@image.ntua.gr

Abstract

In this paper we propose an object detection approach that extracts a limited number of candidate local regions to guide the detection process. The basic idea of the approach is that object location can be determined by clustering points of interest and hierarchically forming candidate regions according to similarity and spatial proximity predicates. Statistical validation shows that the method is robust across a substantial range of content diversity while its response seems to be comparable to other state of the art object detectors.

1. INTRODUCTION

The rapidly expanding research and analysis on object detection and recognition imposes the use of reliable machine vision systems. Recent object detection approaches fuse detection results obtained by robust part detectors, thus eliminating noise artifacts and resolving occlusion phenomena. However, object detection remains a challenging problem due to the wide range of variability in the monitoring conditions as well as to the diversities in scale, location, orientation (up-right, rotated), and pose that real objects adopt.

The goal of an efficient object detector is the accurate determination of objects location, extent and shape. To efficiently approximate these parameters, there are many challenges that should be faced. These are mainly associated with the variability of poses and orientations that the object of interest can adopt. Monitoring parameters such as illumination conditions, camera movements, aliasing and camera specifications also correspond to issues that should be considered

Through this section we provide a brief review of important existing object detection techniques in static images and video. However, a full review of the object detection literature is beyond the scope of this paper. In an effort to distinguish the existing techniques, we can classify them into four broad categories, with respect to the information that they consider. However, some overlapping between categories can occur.

Bottom-Up Feature-Based Approaches aim to find structural features that exist even when the pose, viewpoint, or lighting conditions vary, and then use them in the detection procedure. *Davis & Sharma* [1] proposed a contour-based method to detect humans in widely varying thermal images. At first, regions of interest (ROIs) are determined through statistical background subtraction. Subsequently, gradient information within each region is extracted and used to form a contour saliency map. Morphological operations are also employed to fill broken boundaries. The combination of gradient and contour-based features is an additional class of descriptors that has been extensively used. Their efficacy relies on the robustness of gradient-based descriptors to illumination changes and noise induction. Such features are frequently selected to represent objects boundaries.

The Scale Invariant Feature Transform (SIFT) and shape context have been extensively used for person/face localization [2], [3], [4], [5]. SIFT was initially proposed by *Lowe* [2], [3] and works by combining a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. Geometric histogram [4] and shape context [5] implement the same idea and are very similar to SIFT. In particular, shape context is identical to SIFT descriptors but is based on edges extracted by the *Canny* [6] detector where the location is quantized into nine bins of a log-polar coordinate system.

Another approach also based on gradient orientation distributions is proposed by *Dalal & Triggs* [7]. A dense grid of Histograms of Oriented Gradients (*HoG*) is considered and computed over blocks of specific extent. This representation has proved to be powerful enough to classify humans using a linear SVM. However, it can only process images of limited dimensions and at specific frame rates using a very sparse scanning methodology that evaluates roughly 800 detection windows per image. *Zhu et al.* [8] speed-up *Dalal's* approach by increasing the number of detection windows. Thus, they combine a cascade of rejectors approach with the HoG features. To overcome fast rejection problems, they induce the use of fixed size blocks.

They also use a much larger set of blocks varying in sizes, locations and aspect ratios. The best blocks suited for detection are selected via an AdaBoost relying on a rejector-based cascade. Visual attention has also been used in several object detection approaches in order to account for motion information and conscious search in images. Rapantzikos and Tsapatsoulis [9] have built a robust method of enhancing the accuracy of face detection schemes through a visual attention architecture.

Top-Down Knowledge-Based Methods are rule-based approaches that encode knowledge of what constitutes an object of interest. Knowledge-based methods have the fundamental advantage of attaining to reduce false positive instances of detection by eliminating them through the verification step. Motion characteristics provide important information for both determining regions of interest and assessing whether an object's motion features resemble to human motion [10], [11],[12]. *Viola* and *Jones* [11] have proposed a state of the art human detection approach, which considers prior knowledge on the person's motion and appearance. No separate mechanisms of tracking, segmentation and alignment are supported. The system works by simply selecting the feature set, the scale of the training data and the scales used for detection. The training process uses AdaBoost to select a subset of features and construct the classifier. The classifier consists of a linear combination of the selected features. *Viola* and *Jones* [13] have also proposed a cascade of classifiers architecture to reduce the computational cost.

Template Matching Methods use standard patterns of objects/object parts to describe the object globally or as distinct parts. Correlations between the input image and patterns subsequently computed for detection. *Gavrila* [12] propose a human detection scheme that segments foreground regions and extracts the boundary. Then the algorithm searches for humans in the image by matching edge features to a database of templates of human silhouettes. The matching is realized by computing the average Chamfer distance between the template and the edge map of the target image area. *Wren et al.* [14] describe a top-down person detector based on template-matching. However, this approach requires domain specific scene analysis. *Castillo* and *Chang* use fast template matching as a focus of attention [14] and the algorithm proceeds by discarding locations where there is no silhouette that matches the human body.

In the appearance-based methods, the models (or templates) are learned from a set of training images. These learned models are then used for detection. Appearance-based methods [16] rely on techniques from statistical analysis and machine learning to find the relevant characteristics of images containing objects. The learned characteristics are in the form of distribution models or discriminant functions that are consequently used for detection.

In contrast to all techniques described above, *Integration of Parts detectors* fuses the detection results derived by robust part-based detectors. *Forsyth* and *Fleck* [17] introduced body plans for finding people in general configurations. *Ioffe* and *Forsyth* [18] then assembled body parts with projected classifiers or sampling. However, the aforementioned methods rely on simplistic body part detectors. The employed representation models body parts as bar shaped segments. An improvement on the modeling of body part relations is given by *Mikolajczyk et al.* [19], where local orientation position features are extracted by gradient and Laplacian based filters. The spatial layout of the features, together with their probabilistic co-occurrence captures the appearance of the parts and their distinctiveness. The features with the highest co-occurrence probabilities are learnt using AdaBoost and the detected parts are combined with a joint probabilistic body model. The features deployed in [19] have proven to describe the shape better than prior descriptors. Parts-based object detection and recognition systems were further extended in [20] and [21], where several methodologies have been developed based on either bottom-up or top-down considerations. Bottom-up approaches were used to group together body parts found throughout a sequence while in top-down approaches, human models are built automatically from convenient poses. The system in [20] has been also used to track humans by detecting the learned models in each frame.

The majority of object detection methodologies approach the localization issue through exhaustive search over the image. Thus, false positive artifacts are eliminated at the cost of distorting the object extent and shape. In this paper we try to tackle with the localization problem through considering a set of interest points describing regions that represent transition areas with high probability. These points are subsequently considered to guide the overall detection process. The basic idea is that interest points are located on objects boundaries or on areas of abrupt changes on the background structure.

In practice we determine interest points' locations through Harris corner detector [22] thus enhancing points characterized by high gradient values along all directions. Points grouping (clustering) takes place by measuring distances on low level visual descriptors. The derived clusters are considered as local image structures representing the content of the local neighborhood.

Structuring the paper, section *2* presents a brief overview of the system's functionality while *sub-sections 2.1* through *2.4* analyze the functionality provided by each distinct module. The experimental setup and the results are presented and evaluated in sub-sections *3.1* and *3.2* respectively. System's performance issues are also considered in terms of both statistical and visual analysis**.** Finally, section 4 draws the conclusions and discusses drifts of further research.

2. PROPOSED FRAMEWORK

The proposed methodology approaches the problem of object detection by initially extracting distinctive invariant features from images that are in turn used for region grouping. The features are extracted from local keypoints derived by the Harris corner detector [7] and are associated with point locations that have large gradients in all directions at a predetermined scale. The selected features reflect color and texture information evaluated on local regions and have proven to be invariant across a substantial range of image deformations and illustration conditions. Apart from considering the distributions of color and intensity metrics the employed features also account for spatial dependencies of intensity levels.

After the feature points have been determined they are grouped via an unsupervised clustering approach. An efficient clustering approach should be characterized by accurate discovery of clusters of arbitrary shape (shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc). Moreover, feature points clustering should provide good efficiency when applied on large databases (i.e. on databases of significantly more than just a few thousand objects) while it should also be invariant to prior knowledge on the application data.

In our approach localization of objects initially involves formation of neighborhoods of clusters according to spatial proximity distance. Subsequently, the clusters belonging to each neighborhood are considered in combinations to check the objects occurrence at each specific location. The final step involves the use of a cascade of boosted classifiers.

The overall functionality supported in our method is illustrated in Fig. 1.

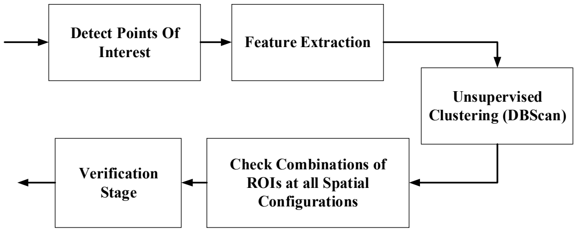
Detect Points Of Interest
Feature Extraction
Unsupervised Clustering (DBScan)
Check Combinations of ROIs at all Spatial Configurations
Verification Stage

Figure 1: Object detection system in diagram form

2.1. Interest Point Extraction and Local Features

Our detection methodology initially considers the point locations extracted by the Harris corner detector [22]. These are determined by evaluating the autocorrelation function of the 2D visual signal (image) within a spatial region of predetermined extent. The autocorrelation function measures the local changes of the patches shifted by a small amount in different directions. Given a shift (Δx, Δy) and a point (x, y), the autocorrelation function is defined as,

$$c(x,y)=\sum_{W}\left[I(x_i,y_i)-I(x_i+\Delta x,y_i+\Delta y)\right]^2 \quad \textbf{(1),}$$

where I denotes the image function and (x_i, y_i) are the points in the window W (Gaussian) centred on (x, y). The shifted image is approximated by a Taylor expansion truncated to the first order terms.

$$I(x_i+\Delta x,y_i+\Delta y)\approx I(x_i,y_i)+\left[I_x(x_i,y_i)\,I_y(x_i,y_i)\right]\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} \quad (2)$$

Substituting (1) into (2) we get,

$$c(x,y)=[\Delta x\ \Delta y]\,C(x,y)\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} \quad (3),$$

where matrix $C(x,y)$ captures the intensity structure of the local neighborhood.

Let λ_1, λ_2 be the eigen-values of matrix $C(x,y)$. The eigen-values form a rotationally invariant description. The presence of a key-point depends on the relation between λ_1 and λ_2. Low levels of λ_1, λ_2 are associated with flat autocorrelation function that corresponds to an image area with no abrupt change in any direction. If only one of the eigen-values is high and the other is low, that implies a ridge shaped auto-correlation function. Such ridges in the auto-correlation function are associated to the occurrence of transition boundaries between different surfaces. High levels of λ_1, λ_2 are reflected to sharp peaks on the autocorrelation function and then shifts in any direction will result in a significant increase that indicates a corner.

Point detection is followed by spatial filtering to eliminate the density of feature points within local neighborhoods. Equation (4) provides a mathematical description of the points filtering operation. Set S corresponds to the set of locations obtained after an initial selection of interest points.

$$\mathrm{S}=\left\{p\in H:\|p-q\|<\delta,\ \forall\, q\in H\right\} \quad (4),$$

where p denotes the point under consideration, H the set of points obtained by the Harris corner detector and δ a parameter denoting the radius of the area being checked. The size of δ is chosen to be significantly smaller than the patch considered for extracting visual descriptors. The underlying idea of selecting the value of δ is to suppress the number of points co-occurring in the same spatial neighborhood while also enabling accurate representation of local image structures.

Keypoints localization is followed by a features extraction procedure. At this step, we consider local square patches of limited extent (significantly larger than δ) and evaluate colour and texture descriptors to represent local areas. Thus, the three most significant components of dominant colour are selected to reflect the local colour content while a statistical descriptor of texture was used to estimate texture variations within images sub-regions. More specifically, at each keypoint location the gray-level co-occurrence matrix is extracted and some features expressing the contrast and homogeneity, within these areas, are evaluated. The co-occurrence matrix displacement vector is selected so as to represent the spatial arrangement of intensity levels while suppressing the induction of intensity peaks associated to noise occurrence.

The texture features considered through this work are the Harralick descriptors of texture and they actually measure the randomness of gray levels distribution. The mathematical expressions of these features are provided through the subsequent equations.

$$Energy : F_1 = \sum_i \sum_j P^2(i,j)$$

$$Entropy : F_2 = -\sum_i \sum_j P(i,j)\log\left(P(i,j)\right)$$

$$Homogeneity : F_3 = \sum_i \sum_j \frac{P(i,j)}{1+|i-j|} \text{ and}$$

$$Contrast : F_4 = -\sum_i \sum_j (i-j)^2 P(i,j),$$

Where $P(i,j)$ denotes the probability of finding pixels with intensity values i, j at spatial arrangements similar to the one defined by the displacement vector.

Feature normalization is also an issue of great concern as it influences the classifiers performance. In this work, L2square-normalization is considered for normalizing the input vectors. The same distance metric is also used by the clustering algorithm [23] for discriminating clusters in the feature space.

2.2. Interest Point Clustering

The clustering approach groups feature points according to similarity and spatial proximity criteria. In practice, the employed algorithm called "Density Based Algorithm for Discovering Clusters in Spatial Databases with Noise" (DBScan) [23] has proven to be powerful enough to discover clusters of arbitrary shape while requiring minimal domain knowledge. The algorithm considers density of feature points as the key aspect for forming meaningful classes in the feature space. Adaptation of the clustering algorithm requires modifying the maximum neighbourhood radius (*Eps*) and the minimum number of points (*MinPts*) in an *Eps*-neighborhood of that point.

The functionality of the DBScan algorithm can be summarized in the following steps.

1. Select arbitrarily a *p* point in the feature space.
2. Select all points that are density reachable from *p* w.r.t *Eps* and *MinPts*.
3. If *p* is a core point, a cluster is formed.
4. If *p* is a border point, no points are density reachable from *p* and DBScan visits the next point of the database.
5. Continue the process until all points have been processed.

Where the terms density reachable and directly-density reachable are associated with the radius of the point's *Eps*-neighborhood and the minimum number of points (*MinPts*) in this neighborhood [23].

After the clusters have been formed, they are examined according to spatial proximity criteria to form spatial neighborhoods in the image plane. The clusters within each neighborhood are subsequently checked in combinations to derive *candidate regions*. The latter are selected to provide local area description and are considered for feature representation. Figure 2 summarizes the operations deployed throughout our methodology:

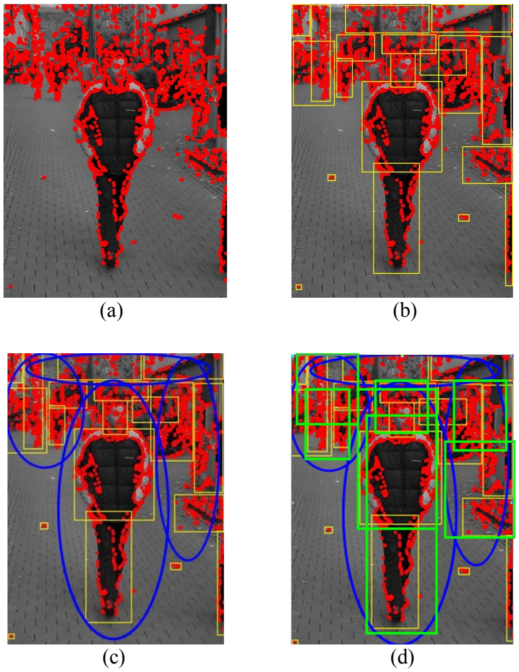
(a) (b) (c) (d)

Figure 2: (a) Points of interest (marked in red), (b) Clusters derived after applying the DBScan (yellow rectangles enclosing interest points), (c) Neighborhoods of clusters on the image, (represented as blue ellipses), (d) candidate regions (enclosed in the light-green rectangles)

As it can be observed the localization process initially involves the determination of neighbourhoods of clusters (represented by blue-coloured ellipses in Fig. 2 (c)) and at the next step, the clusters contained within each neighbourhood are considered in combinations to obtain the candidate regions (light green rectangles in figure 2(d)).

2.3. Feature Extraction

The feature extraction approach is based on evaluating well-normalized local features of image gradient orientations in a dense grid [7]. The basic idea is that local object appearance and shape can be characterized rather well by the distribution of local intensity gradients. The features are selected to be robust to illumination changes and imaging conditions and this is satisfied by encoding the objects boundary orientation and discarding information relative to the local colour or intensity. The histogram representation also enhances the method's robustness to rotation changes.

In practice, the implementation involves dividing the image window into small spatial regions ("*cells*") and for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. More specifically, each pixel calculates a weighted vote for an edge orientation histogram channel based on the orientation of the gradient element centred on it. The votes are evaluated into orientation bins over cells. The combined histogram entries form the representation of possible object boundaries. The vote is a function of the gradient magnitude at the pixel (e.g. $\|\nabla I(x,y)\|$, $\|\nabla I(x,y)\|^2$ or $\sqrt{\|\nabla I(x,y)\|}$). For better invariance to illumination, shadowing, etc., it is also useful to apply contrast normalization of the local features before using them [7]. This can be done by accumulating a measure of local histogram "*energy*" over larger spatial regions ("*blocks*") and using these results to normalize all cells within the block. The histogram of oriented gradients technique has the advantage of capturing characteristic edge or gradient structure information, which can represent local shape with an easily controllable degree of invariance to local geometric and photometric transformations. Thus, translations or rotations make little difference if they are much smaller than the local spatial or orientation bin size. An overview of the feature extraction approach is illustrated in Fig. 3.

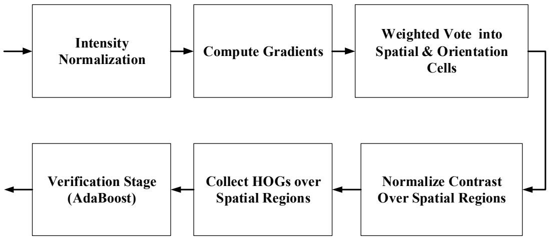
Intensity Normalization
Compute Gradients
Weighted Vote into Spatial & Orientation Cells
Normalize Contrast Over Spatial Regions
Collect HOGs over Spatial Regions
Verification Stage (AdaBoost)

Figure 3: Overview of the feature extraction approach.

2.4. Classification

For the candidate regions defined in 2.2, we evaluate the feature set representing the local image content and employ a cascade of boosted classifiers [13] at the verification stage. The underlying idea in this approach is that smaller and therefore more efficient boosted classifiers can be constructed in order to reject many of the negative areas while detecting almost all positive instances. Simpler classifiers are used to reject the majority of regions of interest before more complex classifiers are called upon to achieve low false positive rates. The overall form of the detection process is that of a degenerate decision tree, also called a "cascade" (see Fig. 4).

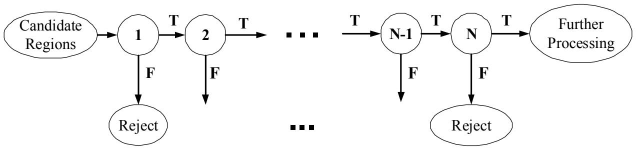

Figure 4: Schematic depiction of the detection cascade. A series of classifiers are applied to regions of interest. The initial classifier eliminates a large number of negative examples with very little processing. Subsequent layers eliminate additional negatives but require additional computation.

3. EXPERIMENTAL RESULTS

3.1 Experimental Setup

Our detection framework is evaluated upon a broad range of images depicting the objects of interest in divert poses and at different illustration conditions so as to resemble the objects occurrence in real scenes. More specifically, the person detection framework considers a part of the PASCAL dataset as training set. The training set contains 2380 positive instances and over 3500 negative examples. Based on this content we have trained three individual classifiers focusing on the full, upper and lower body respectively. However, in the experimental results presented in the following section only the full body classifier is considered so as the results to be comparable with other person detection approaches. The testing set used for evaluating the system's performance contains over 800 images depicting frontal or backward views of human bodies. Regarding the cars and airplanes content, we have produced training sets from the images provided by the PASCAL databases. More specifically, the car detector training set consists of approximately 1800 positive examples and around 2200 negatives while the test set involves 2200 images depicting both negative and positive examples. The corresponding training set for the airplanes contains 1100 positive images and 1300 negative examples.

3.2 Results

The testing framework initially validates the potential of our approach towards the effective determination of the topology and extent of objects of interest. A second objective is to statistically estimate the system's efficiency in detecting objects of interest across a great diversity of monitoring conditions. Figs. 5(a)-(l) depict example of objects detected through our system. Visual inspection of the results indicates that the methodology is quite efficient in determining the objects of interest at their exact location and extent while eliminating the induction of false positives. More specifically, Fig. 5(a)-(d) illustrate the performance in detecting persons at variable scales and under different monitoring conditions. The visual results denote that our method's performance is comparable to other state of the art object detectors. Moreover, its response is closely related to the discriminability provided by the feature set that is used to represent the objects. The robustness against scale changes and its efficacy at suppressing noise artifacts is strongly associated with the image representation.

For the statistical evaluation a number of parameters are being varied and in each case the method's response is evaluated in terms of its potential to detect objects of interest at their actual extent and location.

Through the evaluation procedure, pixels that co-occur in both the detected and the ground truth image are considered as *true positive* (TP) instances of detection.

Similarly, true negative (TN) instances correspond to image locations classified as background by both the detector and the ground truth, while false positive (FP) and false negatives (FN) are determined accordingly. In this paper, we consider the Receiver Operating Characteristic (*ROC*) curves as robust measures for evaluating the algorithmic performance. The ROC curves provide information on the tradeoff between the algorithms specificity (*SP*) and sensitivity (*SE*) [24].

$$SP = \frac{TN}{(TN + FP)} = \frac{TN}{P'} \qquad (6)$$

Similarly the sensitivity value is defined as

$$SE = \frac{TP}{(TP + FN)} = \frac{TP}{P} \qquad (5)$$

The statistical evaluation procedure involves 5-fold cross-validation to guarantee more robust estimation of response. Fig. 6, 7, 8 and 9 provide the statistical curves representing our approach performance. Through the ROC curves, we estimate the trade-off between the systems sensitivity and specificity when adjusting two important parameters of the overall detection methodology. In particular, Fig. 6 illustrates the systems response when the varying parameter is the cell size while Figs. 7 and 8 illustrate performance variations when the DBScan parameters are modified.

A brief study of the derived results illustrated in Fig. 5 indicates that our object detection technique seems to be more robust when detecting cars while the performance of person and airplane detection follows. Particularly, airplane detection seems to have significantly lower performance. This is closely related to the extent of the available training set as well as to the degree to which the training set represents real scenes of airplanes occurrence. The diversity of the testing set is also a significant factor affecting the overall performance.

The effect of DBScan density parameter (Eps) and minimum points (MinPts) on detection performance is illustrated in figures 7 and 8 respectively. As it is observed, the detector seems to provide better results when applied to cars and to airplanes detection problem than humans. This can be explained by considering that people tend to be encountered at more complex scenes. This introduces limitations to the performance of the clustering module, which in turn affects the localization process.

Further comparisons between Fig. 6, 7 and 8 indicate that the variation of the DBScan parameters leads to more abrupt changes on the performance curves than the cell size. This can be explained by considering that the region clustering guides the overall detection procedure. Further comparison between Figs. 7 and 8 indicates that performance is more sensitive to the MinPts parameter than Eps. This can be explained by considering that the minimum number of points affects the overall clustering procedure as it may lead to very large (very low) number of clusters, which in turn obstructs the determination of candidate regions. Experimental and statistical evaluation led to the assessment that the optimum values for the parameters being modified are: *Eps*=0.6, *MinPts*=4 and *cell size*=8x8. These values were also considered in the extraction of the ROC curves.

Finally, Fig. 9 compares the performance of our approach to the approach presented in [7]. The comparison is based on the potential of the two methods to detect persons on the same dataset (INRIA). It can be observed, that our approach attains a slightly better response at the critical area while both approaches present almost the same performance for extreme values of specificity.

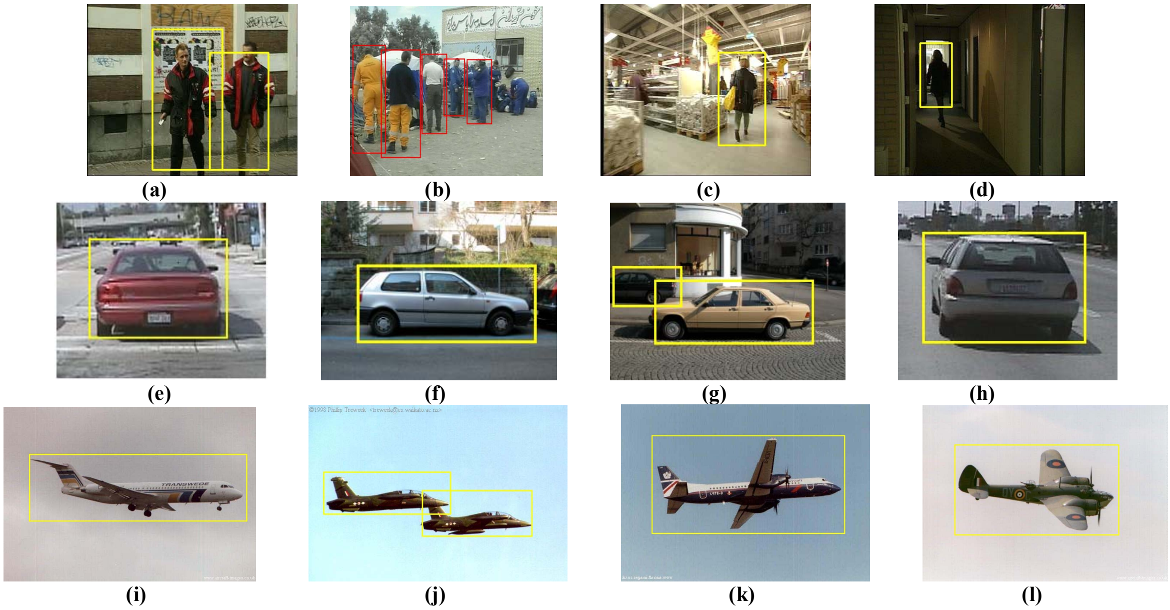

Figure 5: Results of our object detection system (a)-(d): Person detection under different scales and poses, (e)-(h) Detection of cars, (i) – (l): Detection of airplanes**.**

It can also be seen that our detector is invariant to the image size as the overall procedure is guided by the determination of regions of interest.

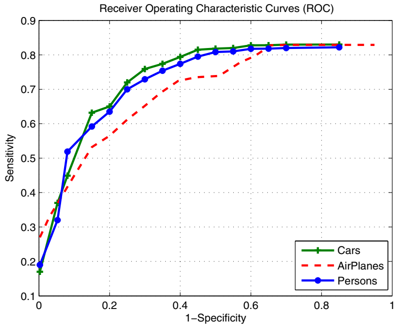

Figure 6: ROC curves obtained by modifying the cell's size.

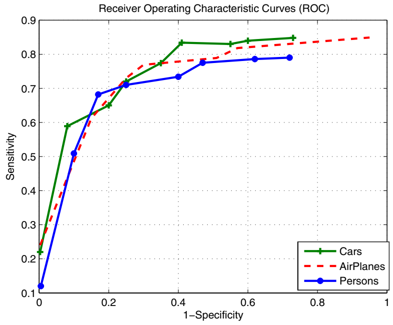

Figure 7: ROC curves derived by modifying the DBScan Eps parameter.

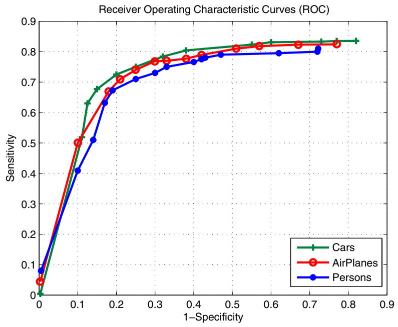

Figure 8: ROC curves derived by modifying the DBScan MinPts parameter.

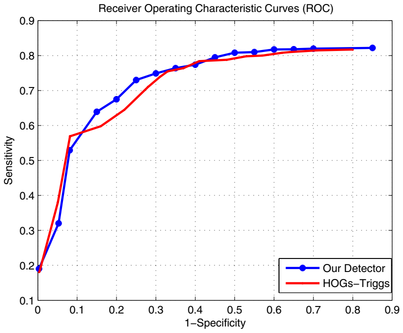

Figure 9: Comparative study of the performance curves representing our method with the approach described in [7].

4. CONCLUSIONS AND FUTURE WORK

Throughout this paper we proposed an object detection approach and evaluated it in real scenes. The approach initially considers points of interest to determine locations close to boundaries. Such points are then grouped according to similarity and spatial proximity criteria. Feature points grouping takes place through an unsupervised clustering approach across a large range of domains (generic detector). The overall localization problem is achieved by forming neighborhoods of clusters. The clusters belonging to each neighborhood are checked in combinations to form candidate regions and in turn to assess the exact location of object's occurrence.

The response is estimated in terms of both visual and statistical evaluation. The test sets are produced so as to depict objects at variable scales and poses. Visual inspection illustrated that the proposed methodology provides accurate determination of the object location and extent. Statistical evaluation also verified that the approach has a performance that is comparable to state of the art object detectors. The overall approach has no limitations related to the image size and the object position.

Although our detection methodology provides quite accurate results, there is still room for further optimization. One of our initial objectives is to evaluate the method's efficiency in terms of detecting objects of interest on other datasets. The TRECVID content is a possible data set for comparisons to other detectors although it does not contain local information in its ground truth. A future direction is the investigation of alternative features (e. g Haar features). The overall detection procedure could also be strengthened by using contour-based information to guide grouping of clusters.

ACKNOWLEDGEMENTS

This research work was supported in part by the European Commission under contracts: FP6-2005-IST-5 IMAGINATION, FP6-027685 MESH and FP6-26978 X-Media.

5. REFERENCES

[1] J.W. Davis, V. Sharma, "Robust Background Subtraction for Person Detection in Thermal Imagery", In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04), vol. 8, pp. 128-132, 2004.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int'l J. Computer Vision, vol. 2, no. 60, pp. 91-110, 2004.

[3] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," Proc. Seventh Int'l Conf. Computer Vision, pp. 1150-1157, 1999.

[4] Ashbrook, N. Thacker, P. Rockett, C. Brown, "Robust Recognition of Scaled Shapes Using Pairwise Geometric

Histograms", In Proc. Of the 6th British Machine Vision Conference, pp. 503-512, 1995.

[5] S. Belongie, J. Malik, J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 4, pp. 509-522, 2002.

[6] J. Canny, "A Computational Approach to Edge Detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, 1986.

[7] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886-893, 2005.

[8] Q. Zhu, S. Avidan, M.C. Yeh, K.T. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", In: Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 1491-1498, 2006.

[9] K. Rapantzikos, N. Tsapatsoulis, "Enhancing the robustness of skin-based face detection schemes through a visual attention architecture", In: Proc of the IEEE International Conference on Image Processing, Genova, Italy, 2005.

[10] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," Computer Vision and Image Understanding, vol. 73, no. 3, pp. 428–440, 1999.

[11] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance", In Proc. of the 9th International Conference on Computer Vision, vol. 1, pp. 734-741, 2003.

[12] D.M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 87–93, 1999.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 511-518, 2001.

[14] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "PFinder: Real-time tracking of the human body", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, 1997.

[15] C. Castillo and C. Chang. "An approach to vision-based person detection in robotic applications", In 2nd Iberian Conference on Pattern Recognition and Image Analysis, vol. 2, pp. 209-216, 2005.

[16] I. Haritaoglu, D. Harwood, and L. Davis,"W4S: A real time system for detecting and tracking people in 2.5D", In Proc. Of the European Conference on Computer Vision (ECCV), pp. 877-892, 1998.

[17] D. Forsyth and M. Fleck, "Body Plans", In Proc. of the International Conference on Computer Vision and Pattern Recognition CVPR, pp. 678-683, 1997.

[18] S. Ioffe and D. Forsyth, "Probabilistic Methods for Finding People", International Journal of Computer Vision, vol. 43, no. 1, pp. 45-68, 2001.

[19] K. Mikolajcyk, C. Schmid, "A performance evaluation of local descriptors", In Proc. Of the International Conference on Computer Vision & Pattern Recognition (CVPR), pp. 257-263, 2003.

[20] D. Ramanan, D.A. Forsyth, A. Zisserman, "Tracking People by Learning their appearance", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 65-81, 2007.

[21] R. Fergus, P. Perona, and P. Zisserman, "Object class recognition by unsupervised scale invariant learning", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264-271, 2003.

[22] C. Harris and M.J. Stephens, "A combined corner and edge detector", In Fourth Alvey Vision Conference, pp. 147–152, 1988.

[23] M. Ester, H. P. Kriegel, J. Sander, X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In Proc. Of the 2nd Int. Conf on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996.

[24] H. Christensen and W. Forstner, "Performance characteristics of vision Algorithms" Machine Vision Applications, vol.9, no. 5-6, pp. 215-218, 1997.