A SALIENCY-BASED RATE CONTROL FOR PEOPLE DETECTION IN VIDEO

*Simone Milani, Riccardo Bernardini, Roberto Rinaldo**

DIEGM - University of Udine
Via Delle Scienze, 208 - 33100 Udine - Italy
e-mail: simone.milani@uniud.it, bernardini@uniud.it, rinaldo@uniud.it

ABSTRACT

Most of latest-generation multimedia systems are equipped with increasingly-effective object detection algorithms (e.g., intelligent video surveillance systems, augmented reality applications, sharing platforms for multimedia data, etc.). Unfortunately, images and video are usually available in compressed formats, which makes object detection more difficult because of the additional distortion noise. In this paper we show that it is possible to mitigate this problem by introducing a rate allocation algorithm that preserves important details for object identification algorithms. We propose a saliency map that identifies crucial elements for detectors. Then, we map saliency values to the value of the quantization parameter to be used by the video coder. Experimental results on HEVC coder show that the proposed rate control algorithm improves the accuracy with respect to the standard strategy.

Index Terms— denoising, object detection, adaptive filtering, saliency map, HOG

1. INTRODUCTION

Nowadays, most video systems employ Artificial Intelligence (AI) solutions to achieve a more precise understanding of the acquired scene. Object detection strategies allow these systems to go beyond the mere appearance of pixels, connecting them with the reality in itself that lies behind each image/video. As a matter of fact, algorithms for the recognition of objects and scenes play a crucial role in several applications like video surveillance systems, augmented reality applications, multimedia sharing platforms (where they are used to enhance classification and retrieval), and many more. Unfortunately, most of the multimedia contents processed by these software modules are available in compressed formats, which permits coding an image or a video with a limited amount of bits at the price of an additional distortion. Data compression is still a need since transmission channels and storage facilities have limited capacities. The additional coding noise reduces the accuracy and the precision of object detection algorithms since it alters the features employed in the classification. Most of the presented algorithms rely on the the statistics of the orientation of edges (usually characterized by Histograms of Oriented Gradients or HOGs) and color histograms [1, 2]. Compression standards typically result in a low-pass transformation which spatially smoothes the original image/video samples and modifies the color information. Moreover, at high compression rates, blocking and ringing artifacts introduce some artificial edge information that is not related to the real scene recorded by the device.

*This work was partially supported by the POR FESR 2007 – 2013, Friuli Venezia Giulia Regional Project "Barcotica."

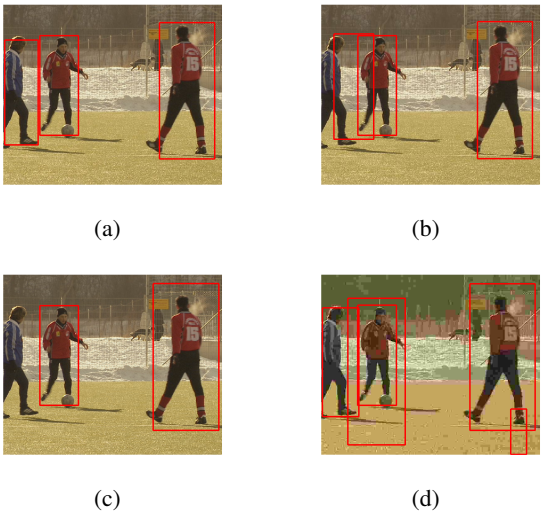


Fig. 1. Performance of object detection with different coding noise levels. The adopted query model is `persons`. The compression ratios are a) 93 % b) 95 % c) 98 % d) 99 %.

Fig. 1 shows the effects of compression noise on the object detection algorithm described in [1] for different compression ratios. It is possible to notice that the higher the compression level (i.e., the coarser quantization operated on the signal) the lower the precision of the algorithm. More precisely, as the quality decreases, the percentage of correct hits decreases since coding artifacts and distortion lead to false hits and misses. Moreover, the object localization becomes coarser preventing a precise selection of the region where the object is present (see Fig. 1.b and Fig. 1.d). As a matter of fact, compression proves to be a delicate task that needs to take into account the effects of the choice of coding parameters on the final performance of the object detection algorithm. Previous works target the problem of coping with noise in object detection by preprocessing the input signal (image or video) with denoising filters. A denoising method for biological macromolecule detection has been proposed in [3] where a set of rotation-equivariant nonlinear filters is employed to denoise contours and perform a rapid object detection in microscopical images. The approach proposed in [4] adopts a noise reduction strategy for pavement images based on wavelet packets. In [5], the authors adopt an adaptive strategy that changes the low-pass behavior of the filter according to the characteristics of the image. However, most of the proposed solutions focus on acquisition noise depending on the capturing conditions and device characteristics, while little work can be found targeting compression. It is possi-

ble to deal with coding distortion by training the detection algorithm using signals in the compressed domain (like [1]). Unfortunately, in practical applications, it is extremely difficult to forecast and model the rate-distortion constraints of the signal since these depend on the characteristics of the signal to be transmitted, the available transmission bandwidth or storage space, the adopted bit allocation strategy [6].

As a matter of fact, compression effects on object detection are to be directly mitigated during the choice of coding parameters. This paper presents a rate control strategy that maximizes the performance of object detection by accurately selecting the most appropriate quantization steps for each image region. The approach relies on the values of a saliency metric [5] that aims at characterizing the crucial features of images for an artificial eye rather than salient points for a human user. This metric is then used to increase or decrease the quantization steps in order to preserve important details while keeping the bit rate within the target limits. The approach has been implemented on the HEVC coder [7] and validated using the object detection algorithm in [1]. Experimental data show that the proposed rate control permits improving the hit/miss statistics reducing the probability of false hits at different bit rate values.

In the following, Section 2 overviews some of the object detection strategies that have been proposed in the literature. Section 3 describes the proposed algorithm, whose test results are reported in Section 5. Conclusions (Section 6) end the paper.

2. OBJECT DETECTION ALGORITHMS

Object detection has been studied for about four decades producing a wide range of object detection techniques. This research effort has been fostered by the many application scenarios where object detection can be employed (e.g., active video surveillance, assisted or autonomous drive, database search, data classification) and by the need of increasing the robustness of existing algorithms to different lighting conditions, poses, scales, locations, and geometries (deformable objects). As a matter of fact, a modern object detection algorithm can be divided into five parts: pre-processing and normalization, local rectification and compensation of small shape variations, computation of descriptor set, machine learning classification, and post-processing to fuse multiple detections.

The pre-processing and rectification phases are very important since they permit compensating light changes and variations in the shape and positions of the object. Usually they include normalizing operations like gamma correction, local contrast normalization, and highlight suppression.

Then, the resulting pixels are processed to generate a set of descriptors for the objects in the image. Different descriptors have been proposed in the literature, like binary patterns from image pixels or the output of a set of steerable filters. The most recent algorithms employ edge orientation histograms. The input image is divided into cells, and for each pixel in the cell the algorithm computes the edge orientation from the image gradient. The histogram of these values is then computed, and the operations are iterated at different image scales. Histograms of gradients are nowadays widely adopted and provide the basis for the popular SIFT [2], HOG [8], and Generalized shape Context methods.

In the classification phase, the descriptors are classified partitioning the feature space into regions. In this case, Support Vector Machine (SVM) classifiers are widely adopted and perform quite well provided that an adequate learning phase is operated. Since the operation is iterated at different scales and multiple detections could emerge from the object detection process, it is necessary to combine and polish the responses of the classifier in order to make the detection more accurate.

All these elements play a significant role in the final accuracy, but in this paper we are going to focus on the first phase. In fact, noise and alterations might severely compromise the efficiency of the algorithm since they significantly alter the generated descriptors. As a result, the experimental data do not fit the partitioned space any longer, and several false positives and negatives can be found in the detection phase. Pre-filtering the image before object detection could be an interesting solution provided that the adopted filters take into consideration the following object detection operations.

In this work, we focus on detection schemes based on histogram of gradients because of their widespread use and the possibilities they offer in implementing real-time object detection systems. Since descriptors are computed from gradient orientation and color information, it is necessary to design an edge-preserving rate allocation strategy in order to prevent the alteration of important features due to coding. In the following section, we present a saliency map that identifies which regions are crucial for object detection and which are not.

3. THE PROPOSED SALIENCY MAP

As previously mentioned, most of the object detection algorithms are based on analyzing the distribution of gradients and colors along object borders. Having this in mind, it is possible to design a saliency metric that combines different features into a normalized value. In particular, the input image is divided into blocks (whose size can be configured according to image resolution), and the included pixels are processed computing three different metrics. In fact, it is possible to observe that object detection strategies are sensitive to edge strength, the stationarity of edge directions, and color contrast along the main orientations of an image. These features can be characterized by a set of parameters that are described in the following subsections.

3.1. Edge strength

In order to characterize the significance of edges in different regions of the image $I(x, y)$, it is possible to compute horizontal and vertical Sobelian gradients (named $S_x(x, y)$ and $S_y(x, y)$, respectively) on the whole image using a 3×3 Sobel operator. Then, these two gradient maps are merged into a common measure of the gradient strength named

$$G(x, y) = round\left(\frac{(|S_x(x, y)| + |S_y(x, y)|)}{16}\right) \qquad (1)$$

where the rounding operation is used to smooth the smallest variations and the normalization factor has been decided after a set of experimental tests. Then, values $G(x, y)$ for the current image are stored in a histogram and the 85-percentile T_g is computed. By selecting those pixel positions (x, y) such that $G(x, y) > T_g$, it is possible to consider only those positions that present relevant gradient information.

3.2. Regularity of edge strength

Re-using the previous results, the orientation of borders $A(x, y)$ can be computed as $A(x, y) = \tan^{-1}(S_y(x, y)/S_x(x, y))$. From these

values, it is possible to compute the local regularity of edges

$$R_A(x,y) = \frac{|A(x,y) - A(x+1,y)| + |A(x,y) - A(x,y+1)|}{2}. \quad (2)$$

For each image block (referenced with the index b), it is possible to compute the variance $\sigma_{R,b}$ of $R_A(x,y)$.

3.3. Contrast along edges

For every pixel position (x,y) such that $G(x,y)$ is greater than T_q, the algorithm evaluates the color differences along the edge orientation, i.e., it computes

$$C(x,y) = \frac{|R(x+\delta_x, y+\delta_y) - R(x-\delta_x, y-\delta_y)|}{3} + \frac{|G(x+\delta_x, y+\delta_y) - G(x-\delta_x, y-\delta_y)|}{3} + \frac{|B(x+\delta_x, y+\delta_y) - B(x-\delta_x, y-\delta_y)|}{3} \quad (3)$$

where $(\delta_x, \delta_y) \in \{-1,0,1\}^2$ is a displacement array aligned along the normal direction to the edge in (x,y). This parameter proves to be extremely important in identifying those objects that can not be easily detected from the background since a low value of $C(x,y)$ denotes a limited contrast around the edge.

For the current pixel block, the proposed saliency metric evaluates the value $\overline{C}_b$, which averages the values $C(x,y)$ for those pixel locations (x,y) in the b-th block such that $G(x,y) > T_g$.

These three metrics are then combined into a single saliency value that is assigned to each block of the image.

3.4. Final saliency metric

The input (uncompressed) image is divided into pixel blocks (indexed with the variable b), and for each of these the algorithm computes $\sigma_{R,b}$ and $\overline{C}_b$.

Starting from this, the saliency value assigned to the current block is

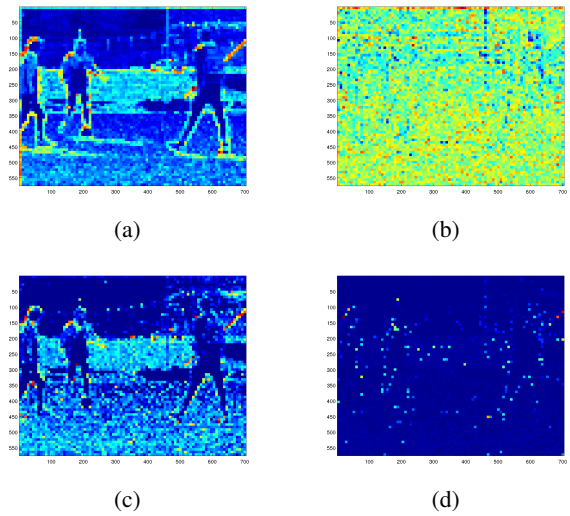
$$S_b = 1 - \begin{cases} \frac{\overline{C}_b}{K} & \text{if } 5 < \sigma_{R,b} < 500 \\ \frac{\overline{C}_b * 10}{K} & \text{otherwise,} \end{cases} \quad (4)$$

where the normalization constant K has been computed in order to make S_b vary in the range $[0,1]$. Parameter K is computed after a set of experimental trials on a set of test sequences. In case $S_b > 1$, its value is clipped as $S_b = 1$, and viceversa, if $S_b < 0$, $S_b = 0$. The value of S_b depends on $\sigma_{R,b}$ since for low and high variance signals, the value of contrast long the edge must be emphasized to fit values within a proper range.

Fig. 2 reports the value of the three parameters G, $\sigma_{R,b}$, $\overline{C}_b$, together with the final saliency value S_b, for the first frame of the sequence `soccer`. It is possible to notice that the highest values for the metric are to be found in proximity of people in the image.

4. RATE CONTROL ALGORITHM

The saliency metric presented in the previous section can be used to select the quantization step in a rate allocation scheme. The main idea is to decrease the quantization parameter (reducing the source coding distortion) as the saliency value increases. In this work, we implemented this strategy considering the emerging video coding standard HEVC [7].



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

Fig. 2. Saliency values for frame 0 from sequence `soccer` (on 8×8 blocks). a) $G(x,y)$ b) $\sigma_{R,b}$ c) $\overline{C}_b$ d) S_b.

The starting architecture is the rate control algorithm implemented in the HEVC reference software HM 7.0, which assigns to the i-th frame to be coded the target bit rate T_i given the available transmission bandwidth, the frame rate, the GOP structure, and the bit allocation statistics of the previous frames. The operations that lead to the computation of the bit budget T_i are the same of the standard rate control algorithm. A quadratic rate-distortion model allows to relate the target bit rate to an average quantization parameter $\overline{QP}_i$ for the current frame. The model is implemented in the original rate control algorithm itself. Then, the standard rate control starts coding the current frame dividing it into Coding Units (CU) and processing each unit independently. The starting QP value is the average parameter $\overline{QP}_i$, but after coding each CU the value is updated according to the remaining bits available for the current frame.

In the proposed strategy, the b-th CU of the i-th frame is coded with the quantization parameter $QP_{i,b}$ which is computed as

$$QP_{i,b} = \begin{cases} \overline{QP}_i - 6 & \text{if } S_b > 0.92 \\ \overline{QP}_i - 3 & \text{if } S_b > 0.8 \\ \overline{QP}_i & \text{if } S_b > 0.73 \\ \overline{QP}_i + 3 & \text{if } S_b > 0.6 \\ \overline{QP}_i + 6 & \text{otherwise.} \end{cases} \quad (5)$$

Note that the higher QP the bigger the quantization step (i.e., the lower the quality). The thresholds have bee computed after an extensive set of experimental tests that optimized the detection capability of the system.

5. EXPERIMENTAL RESULTS

The proposed approach has been tested using the object detector in [1] for people tracking on different video sequences. The adopted system represents objects using mixtures of deformable part models. More precisely, the whole algorithm inherits the key idea of the HOG-based approach [8] but extends object models introducing deformable parts. The HOG-based descriptors are then classified using a latent SVM classifier. Although tests were carried out using this approach only, we believe that the proposed solution can be

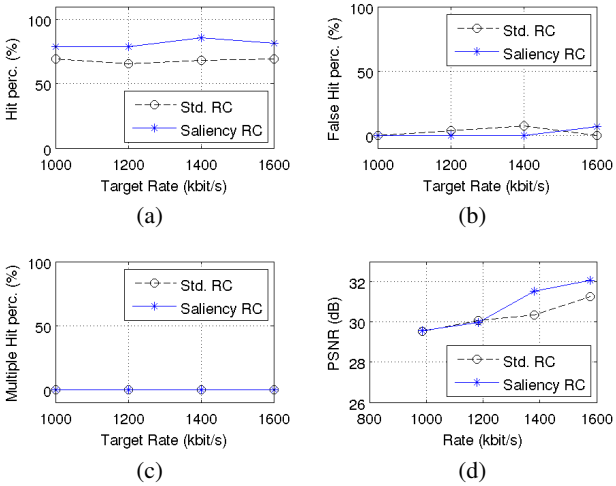


Fig. 3. Performance of different rate control algorithms on the sequence `soccer`. Graphs show the percentages of true hits (a), false hits (b), multiple hits (c) vs. the target bit rate. Rate-distortion performance is reported as well (d).

extended with success to any HOG-based object detector. Person object class was loaded (based on the PASCAL VOC 2006 dataset [9]) and tested on the raw uncompressed video sequence on a set of randomly-selected frames. The outcomes of object detection on the uncompressed sequence will be the ground truth for our evaluation.

Every sequence is then coded with different target bit rates in order to vary the amount of coding distortion introduced in the sequence, and the object detection algorithm is then re-run on the reconstructed video signal (using the same set of frames). The detected objects are then compared with the ground truth data, and two detected objects are matching (hit) in case at least 75% of the pixel coordinates in the bounding boxes are corresponding. The ground truth data are obtained by running the object detection tool on the uncompressed sequence, and checking the feasibility of the classification. From this mapping, we compute the percentage of correct hits (computed from the results on compressed sequence and results on the uncompressed version) and the percentage of false hits, together with the precision of the detection, i.e., the percentage of corresponding pixel positions in the bounding boxes.

Results for the sequence soccer are reported in Fig. 3. The precision obtained by the proposed rate control (in terms of bit rate) is the same of the reference rate control. It is possible to notice that the proposed strategy is able to improve the percentage of correct hits by approximately 20% at all the target bit rates. It is possible to notice that despite at low bit rates the RD-performance of the two algorithms is similar, our approach permits obtaining higher precisions. Thus, image quality for a human user does not necessarily correspond to the quality required by the object detector. Fig. 3 also shows the percentage of false and multiple hits, i.e., the percentage of times that a block is detected more than once. The saliency based algorithm permits reducing the probability of false alarms with respect to a standard approach. It is possible to conclude that the presence of compression noise increases the rate of missing targets, while false alarms keeps low. It is also possible to notice that the precision of the localization is not compromised to much for the `soccer` sequence since the uncertainty on the object localization is quite limited (as shown by Fig. **??**c) Table 1 reports the average performance

Table 1. Performance of saliency maps / standard rate control methods

sequence	True Hits perc. (%)	False Hits perc. (%)	Average PSNR (dB)
harbour	100/100	0/0	30.84/30.70
ice	99/99	12/14	36.82/36.73
crowdrun	51/50	5/6	30.39/30.78

values for sequences `harbour` (coded with rates 2.0, 2.4, 2.8 ,3.2 Mbit/s), `ice` (coded with rates 650, 850, 1050, 1250 kbit/s), and `crowdrun` (coded with rates 3.0, 3.2, 3.4, 3.6 Mbit/s). Data relative to the different bit rates have been averaged for the sake of conciseness. Results show that the proposed solution improves the performance either in the true hits rate or in the minimization of false hits percentage. It is also possible to notice that PSNR quality changes completely independently.

As for the computational complexity, the saliency computation runs in 13 ms on a 3.4 GHz processor on an RGB image with PAL format. As a matter of fact, it supports denoising of video frames at approximately 60 Hz frame rate.

6. CONCLUSION

This paper presented a new rate control algorithm that optimizes the choice of the quantization step in order to maximize the accuracy for object detection modules. The proposed saliency map is based on a set of features which give to the encoder an evaluation of the significance of the different regions in a frame. According to the saliency, the rate control routine increases or decreases the quantization parameter value in order to preserve regions with small contrast from excessive smoothing. The approach performs quite well on HOG-based object detectors permitting sensible performance improvements. The proposed solution proves to be suitable for intelligent video surveillance systems, where both compression and detection accuracy are crucial points. Nevertheless, future research will consider different application scenarios and will test the approach using different object detection algorithms. Moreover, rate control parameters will be optimized within a train-validate-test paradigm involving multiple loops of cross validation to maximize the effectiveness of the final algorithm.

7. REFERENCES

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627 –1645, sept. 2010.

[2] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] M. Reisert and H. Burkhardt, "Equivariant holomorphic filters for contour denoising and rapid object detection," *Image Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 190 –203, Feb. 2008.

[4] Yongxia Zuo, Guoqiang Wang, and Chuncheng Zuo, "Wavelet packet denoising for pavement surface cracks detection," in *Computational Intelligence and Security, 2008. CIS '08. International Conference on*, dec. 2008, vol. 2, pp. 481 –484.

[5] Simone Milani, Riccardo Bernardini, and Roberto Rinaldo, "Adaptive denoising filtering for object detection applications," in *Proc of ICIP 2012*, Oct. 2012.

[6] S. Milani, L. Celetto, and G.A. Mian, "An Accurate Low-Complexity Rate Control Algorithm Based on (ρ, E_q)-Domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 2, pp. 257–262, Feb. 2008.

[7] Philippe Bordes, Gordon Clare, Felix Henry, Mickael Raulet, and Jerome Viéron, "An overview of the emerging hevc standard," in *Proc of ISIVC 2012*, jul. 2012.

[8] Navneet Dalal and William Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, no. 3, pp. 886–893, 2004.

[9] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.