Efficient Scale and Rotation Invariant Object Detection based on HOGs and Evolutionary Optimization Techniques

Stefanos Stefanou and Antonis A. Argyros

Institute of Computer Science, FORTH
and
Brain and Mind Graduate Program, University of Crete

{stevest|argyros}@ics.forth.gr – http://www.ics.forth.gr/cvrl/

Abstract. Object detection and localization in an image can be achieved by representing an object as a Histogram of Oriented Gradients (HOG). HOGs have proven to be robust object descriptors. However, to achieve accurate object localization, one must take a sliding window approach and evaluate the similarity of the descriptor over all possible windows in an image. In case that search should also be scale and rotation invariant, the exhaustive consideration of all possible HOG transformations makes the method impractical due to its computational complexity. In this work, we first propose a variant of an existing rotation invariant HOG-like descriptor. We then formulate object detection and localization as an optimization problem that is solved using the Particle Swarm Optimization (PSO) method. A series of experiments demonstrates that the proposed approach results in very large performance gains without sacrificing object detection and localization accuracy.

1 Introduction

Detecting objects in real-world scenes depends on the availability of local image features and representations that remain largely unaffected by illumination changes, scene clutter and occlusions. A Histogram of Oriented Gradients (HOG) [1] is a descriptor that is computed on a dense grid of uniformly spaced cells and employs overlapping local contrast normalization for improved accuracy. The robustness of HOGs has made them a quite popular image patch/object representation. However, object localization based on HOGs requires the evaluation of the similarity of a reference HOG to the HOG computed in each and every possible placement of a window that slides over the image. Additionally, HOGs are scale and rotation dependent representations. Thus, if one needs to detect and localize objects in a scale and rotation independent way, an explicit and exhaustive consideration of all these search dimensions needs to be performed. This exhaustive search in a multidimensional space becomes computationally prohibitive even for very small image sizes. To overcome this, a variety of methods have emerged [2–4]. Typically, they use heuristics that reduce the number of

HOG similarity evaluations in an image by searching only over a coarse grid of candidate object positions or by using local optimization methods. These methods sacrifice location accuracy to gain speed and, thus, have increased risk of inaccurate localization or even object miss.

In this paper we propose a method to perform accurate object localization in any scale and rotation, avoiding the above drawbacks. We start by proposing a variant of an existing [5], rotationally invariant, HOG-based descriptor. The proposed descriptor relaxes the need of considering rotated versions of it. Furthermore, we formulate object localization as an optimization problem that seeks for the image position and object scale that maximizes the match between the rotationally invariant HOG descriptor and its localization in the image. This optimization problem is solved using the Particle Swarm Optimization (PSO) [6] algorithm. The PSO is a heuristic, evolutionary optimization technique, inspired by search mechanisms employed by certain biological species. Large populations of particles (i.e., candidate solutions) are evolved in iterations called generations to eventually land on the global maximum of the function to be optimized. We demonstrate experimentally that, compared to the sliding window search approach, the proposed approach decreases dramatically the number of descriptor/image similarity evaluations that are needed to localize an object in an image.

2 Related Work

In order to reduce the number of HOG descriptor comparisons required for object localization, many methods have been proposed. Typically, these consist of computing and evaluating the descriptor only over a coarse, limited number of window locations where the object is more likely to be located and over fixed window sizes.

Zhu et al [2] used AdaBoost to select the most relevant windows from an image training set, over 250 random windows per image. In addition, they adopted the integral image representation for a faster formulation of their HOG descriptor variant. This representation of images used in HOG strips the Gaussian mask and trilinear interpolation off the construction of the HOG for each block. In [2], the L2-norm used by Dalal and Triggs [1] is replaced by the L1-norm because it is faster to compute with integral images. Overall, near real-time object localization is obtained but with reduced descriptor robustness. Additionally, the search window locations are heavily depended of the training image set. A similar method [7] uses sparse search at runtime to locate parts of the object in search and then improves the localization by applying a pre-learned Partial Least Squares regression model, followed by an dense search around the approximate locations of the object. Other methods [3, 8–11] employ image pyramids or coarse-to-fine hierarchical schemes. Essentially, detailed searches at higher resolutions are focused on areas where there is evidence for the existence of an object from coarser searches in lower resolutions. This strategy reduces the total number of descriptor evaluations. As an example, Zhang et al [3] applied a

multi-resolution pyramid framework on HOGs to produce better performance over the method of [1]. Interestingly, this work demonstrates that the predefined hierarchy performs better compared to the one that is automatically selected by AdaBoost. The method searches each image at one fourth of the original resolution with a constant window size in a dense pattern, identifying regions of the image not containing the reference object. These regions are then excluded from search in finer resolutions and window grids. The resulting method achieves good localization accuracy and faster execution compared to the original HOG. Still, the method does not consider different object orientations and scales, excluding this way a number of interesting search dimensions. The method proposed by Lampert et al [4] uses a branch-and-bound (B&B) search to find the globally maximal region of the search space - the rectangular bounding box enclosing the target - faster than the exhaustive search. This method reduces the computational complexity from $\mathcal{O}(n^4)$ to $\mathcal{O}(n^2)$ for an arbitrary rectangle bounding box, by trading off accuracy for fast convergence.

The PSO based object detection approach proposed in this work exhibits remarkable performance gains over existing sliding window approaches. At the same time, localization accuracy remain largely unaffected. Due to its nature, PSO provides continuous solutions a fact that is particularly important for estimating the true scale and orientation of an object. As a result, objects are localized in subpixel accuracy and at a fraction of the time needed by the other methods. In addition, PSO search operates without any previous knowledge regarding the possible location of objects and requires the adjustment of only very few parameters.

3 The proposed method

A HOG is not a rotation invariant representation. Therefore, when used in object detection tasks, it can only handle objects that are observed at a certain orientation. To overcome this limitation, a new variant of the HOG descriptor was recently proposed. The so called Rotation-Invariant Fast Feature (RIFF) descriptor [5] is based on a HOG computed at a circular support area and uses an annular binning to achieve orientation invariance. We study the use of a RIFF like descriptor in object detection in conjunction with the Particle Swarm Optimization (PSO) [6]. Scale invariance is not easily achievable through modifications of the HOG descriptor. Instead, the capability for scale invariant search for objects is delegated to the employed optimization technique. Essentially, the detection of a reference object in an image amounts to searching for the image position and object scale that maximizes the match between the rotationally invariant reference object descriptor and the descriptor computed at that image part. The degree of match is quantified by employing the Quadratic-Chi histogram distance [12] between the reference object and the candidate image area. In PSO terms, the Quadratic-Chi histogram distance between a reference HOG and the HOG computed at an image region constitutes the objective function to be minimized.

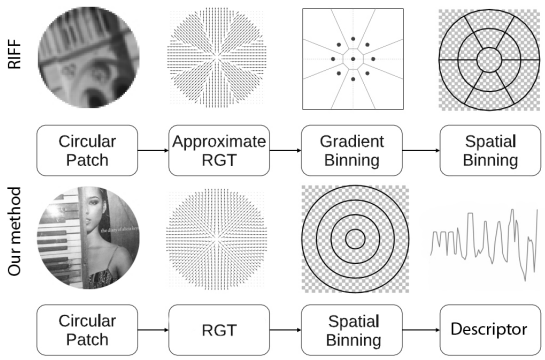

Fig. 1. The RIFF descriptor in comparison with the proposed descriptor variation. The proposed descriptor does not quantize the patch gradients and does not decompose an annulus to sectors.

3.1 Rotation invariance

The original HOG descriptor performs well with objects that are observed at a certain orientation and scale. In order to handle objects that are presented in arbitrary orientations a rotationally invariant object descriptor is required. There are two prominent techniques for achieving rotation invariance. The first [13] treats rotation as a circular shift and uses the magnitude of the Fourier transform, often not sufficiently robust to view point variations. The second [14] uses steerable filters and computes a descriptor for a number of discrete orientations of the filter.

The Rotation Invariant Fast Features (RIFF) descriptor [5] is a recent approach that leverages on the proven methods of SIFT [15] and HOG [1] and provides robustness and rotation invariance. The RIFF descriptor consists of concentric annular cells, applied to image interest points extracted by the FAST [16] detector. Typically, RIFF descriptors consist of four annular cells with the largest diameter being equal to 40 pixels. In each annulus, the image gradient orientations are computed using the centered derivative mask $[-1, 0, 1]$ and rotated to the proper angle according to the Radial Gradient Transform [5] to achieve rotation invariance. The resulting gradients are quantized with respect to their direction for improved performance. Additionally, at each pixel, a local polar reference frame is created for describing the gradient from the radial and tangential directions of the center of the pixel, relative to the center of the descriptor. The coordinates of the gradient in the local frame of reference are invariant to rotation for the given descriptor center. A binning technique is also employed as in CHOG [17]. Computational performance is further improved based on sparse gradient sampling.

In this paper we use a variant of the RIFF descriptor, adapted for whole object recognition. A single circular descriptor is computed that encloses the reference object. The descriptor for a circular image region is computed by first

calculating the edge gradient scale and orientation with a centered derivative mask $[-1, 0, 1]$. We use a signed orientation gradient spanning from -π to π. As in the original RIFF, we define four circular, concentric, non-overlapping annuli. The radii of the circles defining the annuli are computed so that the resulting annuli have the same area. To achieve rotation invariance we rotate the gradients according to the Radial Gradient Transform (RGT) [5] without applying any direction quantization. The final descriptor consists of a histogram of 72 discrete bins (4 annuli $\times$ 18 gradient directions, each). To avoid boundary effects, bilinear interpolation is used to distribute the value of each gradient sample into adjacent histogram bins. Additionally, each pixel's vote in the histogram is weighted by the edge gradient scale. To account for changes in illumination and contrast, a local normalization is performed between cells using the L2-norm followed by clipping the maximum values by a threshold of 0.2 and re-normalizing as in [15]. The final descriptor is the normalized, concatenated rows of the resulting histogram.

3.2 Descriptor distance measure and matching

Since RIFF is a direct representation of a histogram we can use distance measures that are well suited to histogram comparison. We chose the Quadratic-Chi (Q-Chi) histogram distance [12] in order to reduce the effect of differences caused by bins with large values and because of its performance advantages over the simple χ^2 method. According to [12], let P and Q be two non-negative bounded histograms. Let also A be a non-negative symmetric bounded bin-similarity matrix such that each diagonal element is bigger or equal to every other element in its row. Finally, let $0 \leq m < 1$ be a normalization factor. A Quadratic-Chi histogram distance QC between P and Q is defined as

$$QC^{A}{}_{m}(P,Q) = \sqrt{\sum_{ij}\left(\frac{(P_i - Q_i)}{\sum_c (P_c + Q_c){A_{ci}}^{m}}\frac{(P_j - Q_j)}{\sum_c (P_c + Q_c){A_{cj}}^{m}}A_{ij}\right)}. \quad (1)$$

The normalization factor was set to $m = 0.9$.

Concerning descriptor matching, we experimented in comparing the descriptor produced by the original image with the descriptors produced by sub-sampled instances of the same image in different sizes, using the nearest neighbor sampling method. Using the proposed descriptor design, we concluded that descriptors produced from the same image but at different scales typically differ substantially with respect to their Q-Chi distance. More specifically, the nearest neighbor subsampling of an image gave progressively greater distance as the difference in size was increasing. Using bilinear interpolation to match the size of the sub-sampled instances with the size of the original, higher resolution image resulted in much lower influence from scale difference. Finally, using bi-cubic interpolation instead of bi-linear, improves further the results. So, it turns out that it is of importance to match the resolution of the larger image patch by up-sampling the smaller image patch using bi-cubic interpolation prior to computing the descriptor histogram.

3.3 The PSO optimization algorithm

Particle Swarm Optimization (PSO) is an evolutionary technique for the optimization of nonlinear, multidimensional and multimodal functions that is inspired by social interaction. A population of agents, called *particles* is randomly initialized inside the objective function's space. Particles move in search of the function's global maximum for a given number of iterations called *generations*. Each particle is associated with the evaluation of the objective function at its location. Each agent's velocity in the parameter space is determined by three components: a random one, a local one that directs the particle towards its own best position and a global one that directs the particle towards the globally best position. More specifically, the velocity v_i^t for particle i in generation t is given by

$$v_i^{t+1} = K\left(v_i^t + \phi_1 R_1(\boldsymbol{pb}_i^t - \boldsymbol{x}_i^t) + \phi_2 R_2(\boldsymbol{gb}^t - \boldsymbol{x}_i^t)\right), \tag{2}$$

where $\boldsymbol{pb}_i$ is each particle's best position so far, $\boldsymbol{gb}$ is the best position over the whole particles population, $\boldsymbol{x}_i$ the current position of each particle and R_1, R_2 are random numbers in the range [0..1]. Additionally, the so called *constriction factor* K is equal to

$$K = \frac{2}{|2 - \psi - \sqrt{\psi^2 - 4 * \psi}|}, \psi = c_1 + c_2, \tag{3}$$

with $c_1 + c_2 = 4.1$ as suggested in [18]. As the swarm evolves, the agents are expected to locate the global maximum of the objective function and keep oscillating around it. The vast percentage of the computational load of PSO is associated to the evaluation of the objective function for each particle in each generation. Thus, the product of the number of generations to the number of particles is a good indication of the computational load for that PSO parameterization.

Although in principle there are no guarantees for convergence, it has been demonstrated that PSO is able to effectively cope with difficult multidimensional optimization problems in various domains, including computer vision [19].

3.4 Employing PSO for HOG-based object detection

Object detection is formulated as a search task across the three-dimensional parameter space formed by all possible 2D translations and scales at which an object might be present in an image. More specifically, the PSO particles are initialized randomly inside this three-dimensional search space. Each particle corresponds to a single 2D position and scale of the descriptor in the image. The boundaries of this space are determined by the minimum and the maximum scale of the window and the size of the image. To account for partially clipped objects near image borders, the image is padded by mirroring its contents near the edges for 10 pixels. PSO seeks to minimize an objective function which, in our case, is the Q-Chi distance (Eq.1) between the reference object descriptor and and a candidate image window.

4 Experimental results

Several experiments have been performed to assess quantitatively the proposed approach. More specifically, the goal of the experimentation was to evaluate the efficiency and the accuracy of object localization using the proposed technique. The dataset used to evaluate the proposed method is the one used by Tacacs [5] for the evaluation of the RIFF descriptor and consists of images of music CD covers in arbitrary rotation and distance from the camera with partial occlusions and with different backgrounds. The dataset includes 50 different CDs observed in 10 different backgrounds, resulting in a data set of 500 images. For each CD, the clear cover image is provided, based on which the reference descriptor is computed.

In order to evaluate the performance gain of the proposed method, we first produced reference data by locating the position and scale of the window that minimizes the QC distance metric. This was achieved by performing an exhaustive search experiment where all possible object positions and scales were evaluated. To cope with the computational requirements of this exhaustive search experiment, the original images were resized to 320×240 by halving their height and width. With respect to scale, each object has been searched at a minimum window of 60×60 and at a maximum window of 240×240, resulting in 180 different window sizes that were also exhaustively considered.

Next, we ran the proposed method for a variety of PSO parameterizations (number of particles and number of generations). For a particular PSO parameterization our approach reported the position and scale at which an object exists in an image. The localization accuracy of such an experiment was quantified by measuring the F-score (i.e., is the harmonic mean of precision and recall) of the result of our approach and that of the exhaustive experiment. This was repeated for all images. The obtained F-scores for all 500 images are averaged to come up with a single number quantifying the localization accuracy for a certain PSO parameterization. We also measured the accuracy obtained by the sliding window approach where the window displacement step and window size step is equal to $D > 1$ pixels. Essentially, the exhaustive experiment corresponds to $D = 1$ and corresponds to approximately $3,500,000$ objective function evaluations per image. For comparison, running the same experiment with $D = 40$ requires as few as 140 objective function evaluations per image.

Figure 2 summarizes the results obtained from all related experiments. The vertical axis of the plots corresponds to the obtained F-scores. The horizontal axis corresponds to the parameter D. The dashed line corresponds to the average F-score of the sliding window approach as a function of D. As explained earlier, each point in the plot is the average of the F-scores obtained in 500 object searches.

As expected, the exhaustive approach achieves an average F-score of 1 when $D = 1$. As D increases, the average F-score also decreases, reaching the value of 0.25 for $D = 40$. The same plot demonstrates the performance of the proposed approach for a large variety of numbers of particles. For a particular particle count, the number of generations was calculated so that the computational bud-

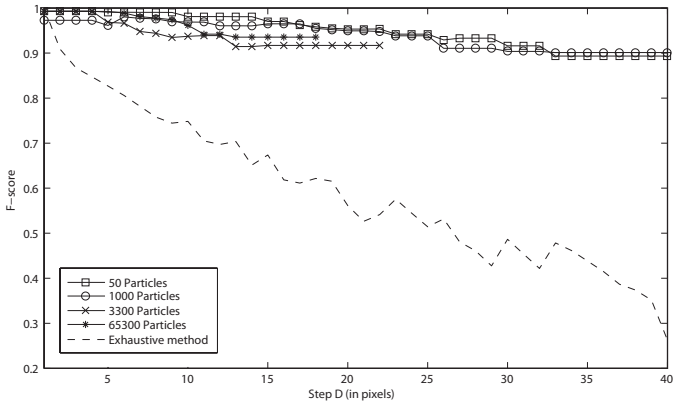

Fig. 2. The mean object localization accuracy with respect to various parameterizations of the proposed algorithm and the exhaustive search approach. The horizontal axis represents the displacement D in location x, y and scale dimensions for the sliding window algorithm. For the proposed method, different plots correspond to different particle numbers. See text for a detailed description.

get required by our method does not exceed the budget of the corresponding exhaustive approach with displacement D. Thus, the intersection of these plots with e vertical line corresponds to algorithms that have the same computational budget and, therefore, require the same execution time. For some particle counts, the curves do not extend up to $D = 40$ because for F values above a threshold, the above mentioned calculation returns zero PSO generations.

As it can easily be verified, the proposed approach keeps an average F-score above 0.9 for all considered computational budgets. This is true even for a budget as low as 140 objective function evaluations ($D = 40$). Thus, when limited computational resources are devoted to object detection, our approach results in more than 3.5-fold improvement in localization accuracy, compared to the sliding window approach.

We also observed that regardless of the parameterization used, PSO is able to localize the object reasonably well in early generations and then performs only minor improvements. Thus, if localization accuracy can be traded with performance, the proposed approach can result in further performance gains.

Another interesting conclusion that can be derived by studying Fig. 2 is that there is no major difference in the use of more generations over more particles. Despite this general conclusion, in individual images and, especially in those that objects appear at smaller scales, it is preferable to use more particles than generations, so that the parameter space is more densely sampled.

Fig. 3. Representative detection results, for objects of arbitrary rotation and scale in different cluttered backgrounds. The proposed approach exhibits robustness to orientation and scale variations as well as to occlusions and illumination artifacts.

Figure 2 presents representative object localization results obtained by the proposed method. In the first four rows, successful detections are shown (in columns 2 to 6) of the reference objects shown in the first column. It can be verified that the proposed approach manages to localize objects despite significant scale and orientation variations as well as partial occlusions, and specular reflections. Interestingly, objects are also accurately localized in the images of the fourth column; these are photos of computer monitors displaying the reference objects. The last row of Fig. 3 shows some of the worst localization results obtained, which we consider as failure cases. In these examples object localization accuracy is small mostly because of the strong specular effects.

5 Discussion

In this paper we formulated object detection as an optimization problem that has been solved with PSO, an evolutionary optimization method. We apply this method to a variant of the HOG descriptor. Experimental results demonstrated that accurate object detection and localization can be achieved at a fraction of the computational cost of the sliding window approach. It is important that PSO has an inherently parallel nature, a fact that can be directly exploited to further reduce the computational time required by employing GPUs. It is also important that the proposed method can very easily be transformed into a tracking framework, which employs object detection at the vicinity of the solution estimated in the previous frame of an image sequence. Current research is considering the

employment of PSO in formulations of object detection problem that exhibit even higher dimensionality.

Acknowledgments

This work was partially supported by the IST-FP7-288146 project HOBBIT.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, San Diego, USA (2005)
2. Zhu, Q., Avidan, S., chen Yeh, M., ting Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: CVPR. (2006) 1491–1498
3. Zhang, W., Zelinsky, G., Samaras, D.: Real-time accurate object detection using multiple resolutions. In: ICCV. (2007)
4. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: CVPR. (2008) 1–8
5. Takacs, G., Chandrasekhar, V., Tsai, S., Chen, D., Grzeszczuk, R., Girod, B.: Unified real-time tracking and recognition with rotation-invariant fast features. In: CVPR. (2010)
6. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: IEEE Int'l Conf. on Neural Networks. (1995) 1942–1948
7. Wu, J., Wei, C., Huang, K., Tan, T.: Partial least squares based subwindow search for pedestrian detection. In Macq, B., Schelkens, P., eds.: ICIP, IEEE (2011) 3565–3568
8. Epshtein, B., Ullman, S.: Feature hierarchies for object classification. In: ICCV, Springer (2005)
9. Agarwal, A., Triggs, B.: Hyperfeatures - multilevel local coding for visual recognition. In: ECCV, Springer (2006) 30–43
10. Amit, Y., Geman, D., Fan, X.: A coarse-to-fine strategy for multi-class shape detection. IEEE Trans. on PAMI **26** (2004) 2004
11. Fleuret, F., Geman, D.: Coarse-to-fine face detection. IJCV **41** (2001) 85–107
12. Pele, O., Werman, M.: The quadratic-chi histogram distance family. In: ECCV. (2010)
13. Kingsbury, N.: Rotation-invariant local feature matching with complex wavelets. In: EUSIPCO. (2006) 4–8
14. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: CVPR. (2008)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004)
16. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: ECCV. (2006) 430–443
17. Chandrasekhar, V., Takacs, G., Chen, D.M., Tsai, S.S., Grzeszczuk, R., Girod, B.: Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In: CVPR. (2009)
18. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evolutionary Computation **6** (2002) 58–73
19. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient Model-based 3D Tracking of Hand Articulations using Kinect. In: BMVC. (2011)