DETECTING SMALL OBJECTS IN HIGH RESOLUTION IMAGES WITH INTEGRAL FISHER SCORE

Roberto Leyva [†], *Victor Sanchez* [†] *and Chang-Tsun Li* [†,‡]

[†] University of Warwick, UK, and [‡] Charles Sturt University, Australia

ABSTRACT

Nowadays, big imaging data are very common in many fields of study. As a result, detecting small objects in very large images is challenging and computationally demanding. Taking advantage of the intrinsic cumulative properties of the Fisher Score, we propose the Integral Fisher Score (IFS) for low-complexity and accurate object detection in big imaging data. The IFS, which is a multi-dimensional extension of the Integral Image, allows computing the Fisher Vector associated with a spatial region using only four operations. This considerably reduces the computational cost of searching for a small query object on a very large target image. Evaluations for the detection of small object on high-resolution HUB telescope and digital pathology images show that IFS attains a high accuracy with short processing times.

Index Terms— Big data, Fisher Vectors, object detection, integral image, local features

1. INTRODUCTION

Large images are nowadays very more common in many fields, e.g., medicine, digital microscopy and astronomy. Consequently, many basic computer vision tasks, such as object detection, now require a significant computational effort to deal with these big data [1]. Object detection is a relatively fast and simple task if the target image is small, multiple copies of the query object do not appear in the target image, and other objects are concisely dissimilar to the query object. However, this task becomes very computationally complex and challenging if the target image is very large and depicts several objects that are similar to the query one, or the query object is very small compared to the target image. To improve detection accuracy, several efficient methods based on local features [2], e.g., SIFT, and neural networks have been proposed [3]. Methods based on local features require only an example of the query object, which is detected by matching its features with those extracted from the target image. Therefore, these methods can detect any object with no *a priori* training process. On the other hand, methods based on neural networks require several examples of the query object(s) for training [4]. Thus, they are restricted to detecting the object(s) provided in the training process.

Despite their effectiveness, the accuracy of the methods based on local features heavily depends on how exhaustive the search for similar features is. An exhaustive search provides a very high accuracy at the expense of long computational times. This becomes an important issue when dealing with very large target images and small query objects, as multiple overlapping regions of different sizes must be analyzed to accurately detect the object. In this paper, we propose the Integral Fisher Score (IFS) to solve this issue. We concentrate on local features due to the flexibility they afford to detect any query object with no training process. Our proposed IFS, which extends the concept of Integral Image, allows computing the Fisher Vector (FV) associated with a spatial region with a very low computational complexity. This allows accurately detecting small objects in big imaging data by searching over several multi-scale regions very fast. Evaluations on high-resolution HUB telescope and digital pathology images with very small query objects show that our proposed IFS considerably reduces computational demands compared to the state-of-the-art methods, while providing a high detection accuracy.

The rest of this paper is organized as follows. Section 2 presents a brief review of the relevant previous work. Section 3 details the proposed IFS. Section 4 presents and discusses the evaluation results. Finally, section 5 concludes this paper.

2. PREVIOUS WORK

Local features have been shown to significantly improve accuracy for image categorization [5] and object detection [6, 7] tasks. For the case of object detection, FVs have further improved the detection accuracy. [8, 9, 10, 11]. A FV is a high dimensional vector that represents a sub-set of features, q, by concatenating their gradient projection over a model, $\mathcal{M}$, that represents the distribution of a feature set, Q, with $q \subset Q$ [8, 12]. In object detection, Q is extracted from the target image, q is extracted from a region of the target image where the query object may be present, and $\mathcal{M}$ is usually a Gaussian Mixture Model (GMM). Despite their ability to improve detection accuracy, FVs are very computationally complex [13, 14], particularly when dealing with big imaging data and small query objects [15]. The are two main ways to address this problem at the expense of sacrificing the accuracy. The

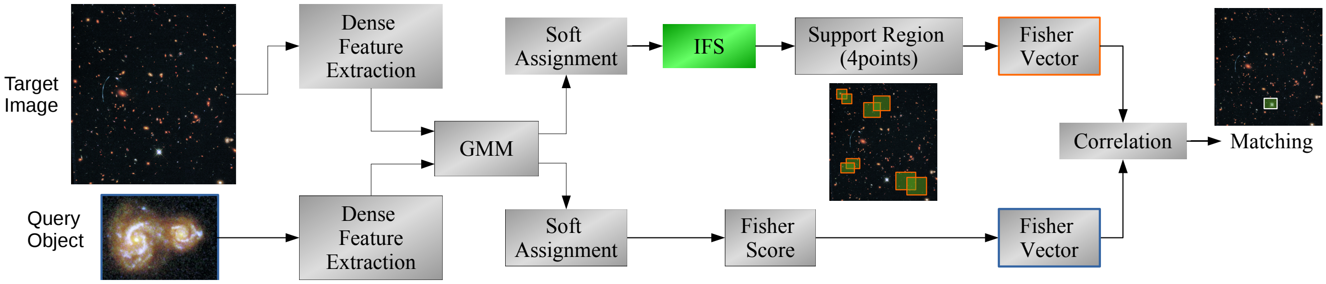

Fig. 1. Object detection. The GMM of densely extracted features of the target image is used for soft assignment. The IFS generates a FV for each support region. The correlation between each FV of the target image and that of the query object is computed to detect the object.

first one involves employing parameter constraint techniques on $\mathcal{M}$; e.g., constraining the number of components of the GMM, the number of iterations to fit the GMM, or the size of Q used to fit the GMM [16, 17, 13]. The second one involves using a sub-optimal $\mathcal{M}$ but with more descriptive features, e.g., by adding color information [18], using improved gradient sources, or hybrid CNN features [19, 14].

Recently, the work in [20] has proposed using a saliency detector as a first step to detect potential areas where the query object may be found. This method only computes the FVs of the salient regions, thus reducing processing times. A similar idea is proposed in [21], which computes FVs from superpixels computed for specific areas.

3. SMALL OBJECT DETECTION WITH IFS

Fig. 1 depicts the framework for small object detection in big imaging data using our proposed IFS. The framework densely extracts features from the target image, whose distribution is fitted to a GMM. The GMM provides the parameters for soft assignment and to map every feature into a high dimensional space. The framework then calculates the IFS from the projected features. Using a pyramid of scales to define several support regions in an overlapping fashion, it computes the FV of each support region using the corresponding IFS – this requires only four operations. Finally, the correlation between each FV of the target image and the FV of the query object is measured to find the exact location of the object.

Given an N-component GMM representing the distribution of the features extracted from the target image, our IFS projects each feature, z_m, via soft assignment onto the n^{th} distribution component:

$$\gamma_m(n) = \frac{w_n p_n\left(z_m|\theta\right)}{\sum\limits_{1\leqslant j\leqslant N} w_j p_j\left(z_m|\theta\right)}, \tag{1}$$

where p_n and $\theta = \{w_n, \mu_n, \sigma_n\}$ are the posterior and parameter set (weight, mean, standard deviation) of the n^{th} component, respectively. In the traditional Fisher Score (FS), the gradient vector that represents the FS of a feature set $Z = \{z_1, \ldots, z_m \ldots, z_M\}$ is given by the concatenation of the corresponding $\mathcal{G}_n^Z$ vectors, $\mathcal{G}^Z = [\mathcal{G}_1^Z \quad \mathcal{G}_2^Z \quad \mathcal{G}_3^Z \quad \ldots \quad \mathcal{G}_N^Z]$, where $\mathcal{G}_n^Z$, corresponding to the n^{th} GMM component is:

$$\mathcal{G}_n^Z = \frac{1}{\sqrt{w_n}} \sum_{1\leqslant m\leqslant M} \mathcal{G}_n^{z_m}, \tag{2a}$$

$$\mathcal{G}_n^{z_m} = \gamma_m(n)\left(\frac{z_m - \mu_n}{\sigma_n}\right). \tag{2b}$$

Note that Eq. 2a sums M features in an unordered manner. We can easily enclose these M features in a spatial region, s, defined by $x_0 \leqslant x \leqslant x_1$ and $y_0 \leqslant y \leqslant y_1$ within an $\{x, y\}$ plane as Fig. 2a illustrates. Therefore, Eq. 2a can be re-written as follows:

$$\mathcal{G}_n^Z = \frac{1}{\sqrt{w_n}} \sum_{z_m \in s} \gamma_m(n)\left(\frac{z_m - \mu_n}{\sigma_n}\right). \tag{3}$$

It is interesting to note that Eq. 3 is very similar to computing the Integral Image of a spatial region s. Let us recall that the Integral Image J_p up to point $p = \{x_p, y_p\}$ within an image I is given by:

$$J_p = \sum_{\substack{1\leqslant x\leqslant x_p\\ 1\leqslant y\leqslant y_p}} I(x, y). \tag{4}$$

To compute the Integral Image of an area s defined by four points, $\{a, b, c, d\}$ (see Fig. 2b) one can simply compute:

$$\sum_{\substack{x_a\leqslant x\leqslant x_d\\ y_a\leqslant y\leqslant y_d}} I(x, y) = J_a + J_d - (J_b + J_c), \tag{5}$$

where $x_a = x_c, x_b = x_d$ and $y_a = y_b, y_c = y_d$. Using the Integral Image generated by the set of $\mathcal{G}_n^{z_m}$ values (see Fig. 2a), Eq. 3 can be computed with four operations:

$$\hat{\mathcal{G}}_n^Z = \mathcal{J}_{a,n}^Z + \mathcal{J}_{d,n}^Z - \left(\mathcal{J}_{c,n}^Z + \mathcal{J}_{b,n}^Z\right), \tag{6}$$

$$\mathcal{J}_{p,n}^Z = \frac{1}{\sqrt{w_n}} \sum_{z_m \in p} \gamma_m(n)\left(\frac{z_m - \mu_n}{\sigma_n}\right), \tag{7}$$

where $\hat{\cdot}$ specifies that the Integral Image is used to compute $\mathcal{G}_n^Z$, and p is the spatial region defined by $1 \leqslant x \leqslant x_p$ and $1 \leqslant y \leqslant y_p$. The proposed IFS for region s is then:

$$\mathcal{G}_s^Z = [\hat{\mathcal{G}}_1^Z \quad \hat{\mathcal{G}}_2^Z \quad \hat{\mathcal{G}}_3^Z \quad \ldots \quad \hat{\mathcal{G}}_N^Z]. \tag{8}$$

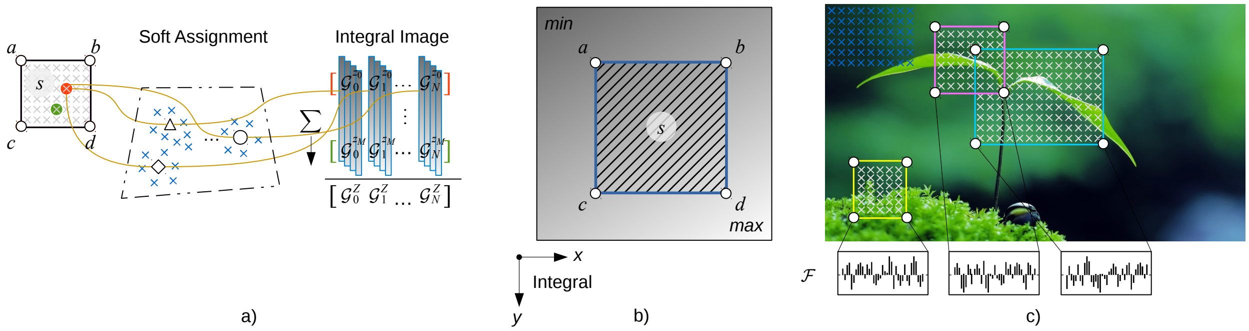
Soft Assignment
Integral Image
a
b
c
d
s
$\sum$
$[\mathcal{G}_0^{z_0}\ \mathcal{G}_1^{z_0} \cdots \mathcal{G}_N^{z_0}]$
$[\mathcal{G}_0^{z_M}\ \mathcal{G}_1^{z_M} \cdots \mathcal{G}_N^{z_M}]$
$[\mathcal{G}_0^{Z}\ \mathcal{G}_1^{Z} \ldots \mathcal{G}_N^{Z}]$
a)
min
max
x
y
Integral
b)
$\mathcal{F}$
c)

Fig. 2. **a)** IFS: features in region s are projected and summed using the Integral Image of each dimension of the projection. **b)** Region s whose Integral Image is computed using the Integral Image values of $\{a, b, c, d\}$. **c)** FVs computed in a multi-scale overlapping fashion.

After computing the IFS, the corresponding FV, $\mathcal{F}_s$, is [12]:

$$\mathcal{F}_s = \text{sign}\left(f_s\right)\left|f_s\right|^{\alpha}, \tag{9a}$$

$$f_s = \sqrt{\mathcal{G}_s^Z \left(\mathcal{G}_s^Z\right)^T}. \tag{9b}$$

where α is a constant used for power-ℓ_2-normalisation. The correlation, ρ_s, between the query object's FV, $\mathcal{F}_q$, and each $\mathcal{F}_s$ extracted from the target image (see Fig. 2c) indicates that the object is present when $\rho_s = 1$:

$$\rho_s = \frac{\mathcal{F}_q \mathcal{F}_s^T}{\|\mathcal{F}_q\| \|\mathcal{F}_s\|}. \tag{10}$$

It is important to note some key differences between our IFS and the work in [22], which also employs the Integral Image. The work in [22] assigns each feature an index pointing to the GMM component that best models the feature. It then uses these indices to compute an Integral Image. Our IFS, on the other hand, uses the features' high dimensional projections to compute Integral Images, one for each dimension of the projection. This allows our IFS to exploit the full power of FVs when working with features.

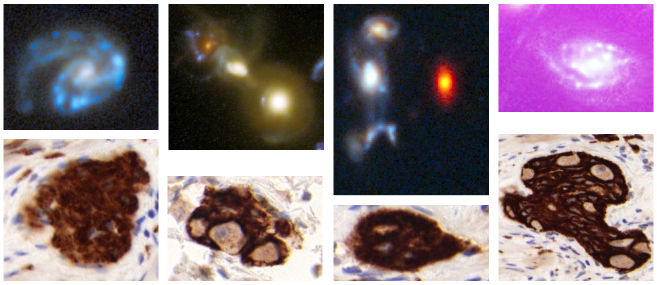

Fig. 3. Example query objects detected in the HUB telescope (1st row-galaxies/stars) and digital pathology (2nd row-cells) images.

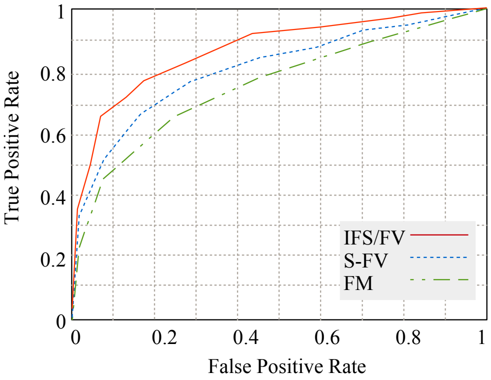

Fig. 4. ROC curve of evaluated methods.

4. PERFORMANCE EVALUATION

Our evaluations use 22 high-resolution HUB telescope [23] and digital pathology images [24] with sizes ranging from 1100×1100 to $27k \times 30k$ pixels for the detection of very small query objects representing 0.005%–2.1% of the target image's size (see Fig. 3). We compare the following four object detection methods. 1. IFS: our IFS with SIFT features (Fig. 1). 2. FV: traditional FS with SIFT features [8]. 3. S-FV: saliency + traditional FS with SIFT features [20]. 4. FM: feature matching (SIFT) with brute force [2]. For the first three methods, a detected query object whose detected region covers at least 80% of object's spatial extent is a true positive (TP), otherwise is a false positive (FP). If the query object is not present in the target image and the maximum correlation, ρ_s, is below a threshold, the outcome is a true negative (TN), otherwise is a false negative (FN). For FM, if at least 50% of the matched features correspond to the query object, then the object is correctly identified (TP). If less than 50% of the query object's features are matched in the target image and the object is not present, the outcome is a TN. Fig. 4 shows the accuracy of the evaluated methods in terms of the Receiver Operating Characteristic (ROC) curve. Table 1 summarises the results from the ROC curve using the Equal Error Rate (EER), and tabulates the average processing time per image. Experiments are run on a Intel i5 CPU with 16 GB of RAM. We observe significant improvements using the IFS over S-FV and FM. As expected, IFS and FV attain the same detection accuracy, as both are equivalent feature representations and compute the same number of FVs for the same regions (at the same number of scales) of the target image. The main difference between them is the computational times. IFS re-

Table 1. Performance of evaluated methods.

Method	Average processing time per image (sec.)	EER
IFS	224.6	**0.198**
FV	1075.0	**0.198**
S-FV	10.5	0.278
FM	22.1	0.321

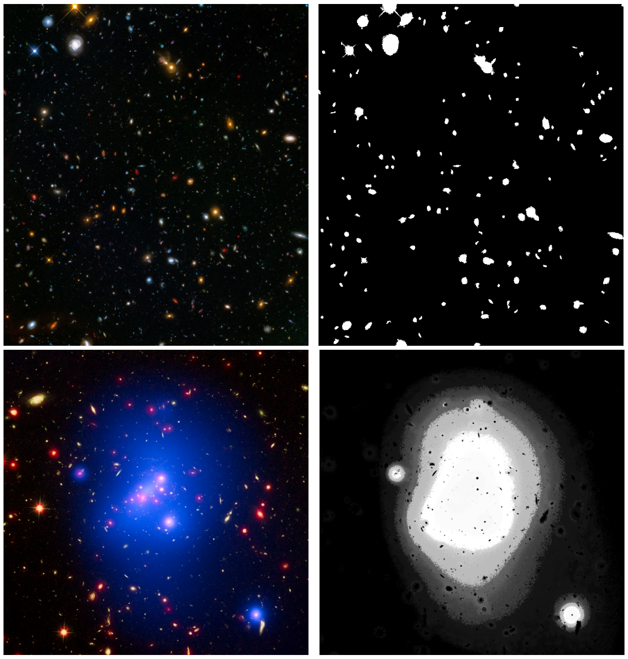

Fig. 5. Saliency detector probability map (right column) for two target mages with different light variations (left column).

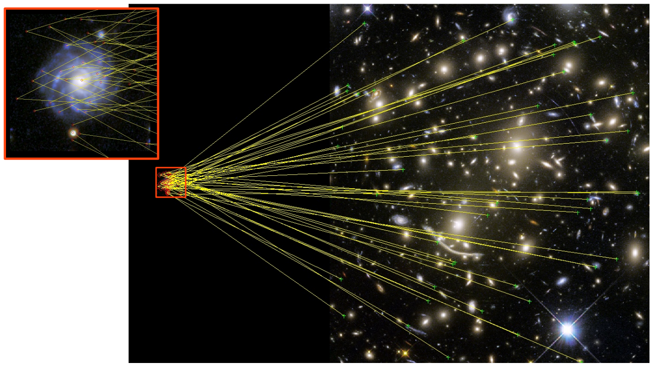

Fig. 6. FM: features of the query object are matched individually all over the target image.

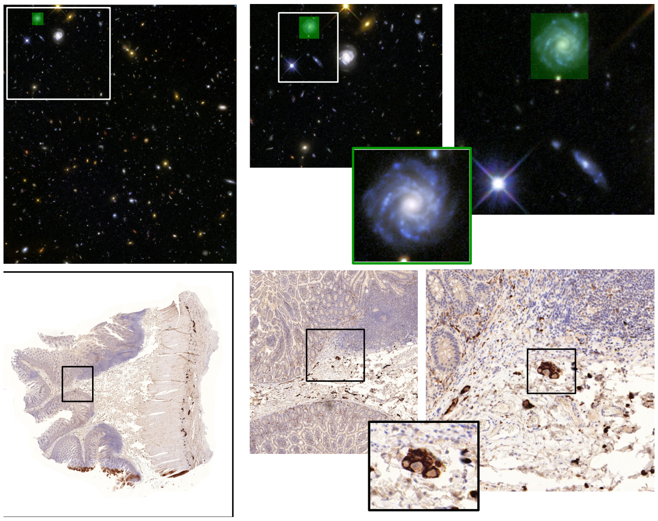

Fig. 7. Example detections of IFS. The query object represents 1.2% (top) and 0.005% (bottom) of target image's size.

duces c.a. five times the computational cost of FV. IFS has a complexity of $O(4SK)$ for S regions and K scales, while the traditional FS has a complexity of $O(MSK)$ for M features. We use $K = 5$ scales for $S = \{0.5, 1, 1.25, 1.5, 2\}$ times the query object's size.

Note that S-FV attains the shortest average processing time, but at the expense of a low detection accuracy. S-FV calculates the FVs only for the areas retrieved by the saliency detector, therefore reducing the number of computations. However, if the saliency detector is not robust to light variations and noise, the accuracy is severely hindered. Fig. 5 illustrates this problem: in the first row, the target image has no significant light variations. Therefore, the saliency detector successfully detects a limited number of regions where the query object may be present. In the second row, the target image has significant light variations. The saliency detector consequently retrieves three large overlapping regions missing completely the query object. FM attains the worst detection accuracy. This method matches features individually and not as a whole for a specific region. Therefore, when searching for a very small object on a large target image with very similar objects, the features from the query object may be matched all over the target image, failing to correctly detect the specific region where the query object is. This issue is illustrated in Fig. 6. IFS, on the contrary, does not fail when the target image depicts several objects that are very similar to the query object, as it matches feature compositions corresponding to local regions and not individual features. Fig. 7 shows successful example detections of IFS. Overall, our IFS maintains the high detection accuracy of FV, while significantly reducing computational times.

5. CONCLUSIONS

This paper proposed the Integral Fisher Score (IFS) for accurate and low-complexity detection of very small objects on high-resolution images. Our IFS is a new way to compute Fisher Vectors that significantly reduces the number of projected features needed for calculation. This allows searching for an object by analyzing several overlapping regions at multiple scales in a fast manner. Evaluations on challenging HUB telescope and digital pathology images show that IFS can significantly speed up the detection process of very small objects while attaining a high accuracy.

6. ACKNOWLEDGEMENTS

This work was funded by the Mexican Ministry of Education - CONACyT and the EU Horizon 2020 project, entitled Computer Vision Enabled Multimedia Forensics and People Identification (acronym: IDENTITY, Project ID: 690907)

7. REFERENCES

[1] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. S. Pedersen, and C. Igel, "Big universe, big data: Machine learning and image analysis for astronomy," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 16–22, Mar 2017.

[2] P. Azad, T. Asfour, and R. Dillmann, "Combining harris interest points and the sift descriptor for fast scale-invariant object recognition," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 4275–4280.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 580–587.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, vol. 2, pp. 2169–2178.

[6] Yang Wang and Greg Mori, "A discriminative latent model of image region and object tag correspondence," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., pp. 2397–2405. Curran Associates, Inc., 2010.

[7] Mario Fritz, Gary Bradski, Sergey Karayev, Trevor Darrell, and Michael J. Black, "An additive latent feature model for transparent object recognition," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., pp. 558–566. Curran Associates, Inc., 2009.

[8] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, Dec 2013.

[9] Basura Fernando, Elisa Fromont, and Tinne Tuytelaars, "Mining mid-level features for image classification," *International Journal of Computer Vision*, vol. 108, no. 3, pp. 186–203, Jul 2014.

[10] R. G. Cinbis, J. Verbeek, and C. Schmid, "Segmentation driven object detection with fisher vectors," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 2968–2975.

[11] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[12] Gabriela Csurka and Florent Perronnin, *Fisher Vectors: Beyond Bag-of-Visual-Words Image Representations*, pp. 28–42, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[13] Y. Kawano and K. Yanai, "Rapid mobile object recognition using fisher vector," in *2013 2nd IAPR Asian Conference on Pattern Recognition*, Nov 2013, pp. 476–480.

[14] Lingqiao Liu, Chunhua Shen, Lei Wang, Anton van den Hengel, and Chao Wang, "Encoding high dimensional local features by sparse coding based fisher vectors," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 1143–1151. Curran Associates, Inc., 2014.

[15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep fisher networks for large-scale image classification," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 163–171. Curran Associates, Inc., 2013.

[16] M. S. Biagio, L. Bazzani, M. Cristani, and V. Murino, "Weighted bag of visual words for object recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 2734–2738.

[17] C. Baecchi, F. Turchini, L. Seidenari, A. D. Bagdanov, and A. D. Bimbo, "Fisher vectors over random density forests for object recognition," in *2014 22nd International Conference on Pattern Recognition*, Aug 2014, pp. 4328–4333.

[18] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, May 2016.

[19] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, Sep 2014.

[20] A. Jose and I. Heisterklaus, "Bag of fisher vectors representation of images by saliency-based spatial partitioning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 1762–1766.

[21] I. Buzcu and A. A. Alatan, "Fisher-selective search for object detection," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 3633–3637.

[22] K. E. A. v. d. Sande, C. G. M. Snoek, and A. W. M. Smeulders, "Fisher and vlad with flair," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2377–2384.

[23] "Hubblesite," `http://hubblesite.org/images/gallery`.

[24] "The cancer genome atlas, national cancer institute, national institute of health.," `https://cancergenome.nih.gov/`.