Overcomplete Representation Acceleration Methods in Source Localization and Object Detection

Zhengye Yang
Electrical, Computer, and Systems Engineering, *Rensselaer Polytechnic Institute*

***Abstract*—Finding a good representation is key to almost all signal processing applications. Overcomplete representation possesses many characteristics that are beneficial to signal processing applications. However, it implicitly brings a heavy computational complexity that impedes tractability in real-world applications. In this paper, we synthesize two publications from image object detection and sensor array source localization that are related to overcomplete representations and corresponding computation acceleration. we also discuss the potential research directions for future work.**

***Index Terms*—feature representation, source localization, direction-of-arrival estimation, object detection**

I. Introduction

Feature representation is the corner stone of many signal processing applications including image classification, object detection, source localization, etc. [10], [13], [18] Yet, designing task-related informative features and building fast algorithms based on those features bring huge challenges. Due to the unique characteristics of each downstream signal processing applications, each research direction has become its own root. Specifically, this paper mainly focuses on synthesising the assigned articles from two different branches: object detection [3] and source localization [19], each direction is presented as an independent subsection in the following sections.

Source localization aims to localize sources using sensor arrays. The sensor array is usually far away from the sources, which makes the distance estimation impossible. Therefore, the task becomes to estimate the direction-of-arrival of sources. One of the primary goal of source localization is to accurately localize sources, which requires the ability to differentiate sources even when sources are closely spaced. Beamforming [12], Capon's method [1] and MUSIC [23] are some of the most well known nonparametric methods. The basic assumption of beamforming is the received signal energy achieves its maximum when steering the sensor array to the source direction. This vanilla method performs well in the simple lab experiment when the sound sources are uncorrelated and well spaced with little white noise. But it suffers from Raleigh limitation that it is not able to differentiate two sources when they are closely spaced. Capon's method (Minimum Variance Distortionless Response (MVDR) filter in audio processing literature) tends to augment the noise toleration ability by minimizing the noise power with a fixed gain of the actual source direction. MUSIC algorithm relies on the spectral decomposition to decompose the received signal to orthogonal signal and noise subspace. This method implicitly needs the sources to be uncorrelated and also a large number of snapshots to ensure its performance. Parametric methods like deterministic maximum likelihood (DML) and stochastic maximum likelihood (SML) perform well with hard-to-acquire accurate initialization [13]. How to accurately localize correlated, close spaced sources under considerable noise within acceptable time span is desired to be explored.

Localizing and recognizing multiple objects within the frame is the key objective of object detection. Both localization and classification branches require good feature representation. There are a significant amount of efforts for extracting useful features from images [11], [14], [15], [18]. Consider the variation of object scales, building object detection algorithms with multiscale feature representation can achieve good overall detection performance. Feature/image pyramid is almost a default choice in designing object detection algorithms [5], [9], [10], [17], [20]. Famous Integeral Chanel Feature (ICF) [5], Deformable Part models (DPM) [10] work well in both pedestrian detection (Caltech [6], INRIA [2], ETH [8] datasets) and general object detection (PASCAL VOC dataset). But the corresponding computational cost also impedes them to be applied in the real-world scenario. There are a lot of ad-hoc efforts for speeding up the specific detection algorithms (e.g. cascaded and coarse-to-fine DPM [9], [20]), and these acceleration methods are all based on precomputed image features. While

constructing feature pyramid is computational-expensive, and the investigation of accelerating feature computation is missing. Finding methods to accelerate the process is key to real-world applications tractability.

Overcomplete representation has been an active research area for many years [7], [16]. Overcomplete representation means the basis vector outnumbers the input signal dimensions. It often comes with robustness to noise, flexibility to match the data structure and easiness to utilize sparsity. Different advantages of overcomplete representation have been utilized in both areas. For source localization, it perfectly aligns the central assumption that there are only relatively small number of point sources in the localization scheme. Sparsity is easily introduced by overcomplete representation. In the vision realm, both primate and computer visual systems benefit from using overcomplete representation to extract visual features, which provide key information for algorithm to detect object within the frame [11], [14], [15], [18]. By utilizing the overcomplete representation, the visual features are more robust to viewpoint changes, lighting and image deformation.

In the following sections, we discuss the problem formulation with overcomplete representation of source localization and image feature representation, the proposed methods from [3], [19] to accelerate computation of such representation and the achieved results with discussion.

II. Motivation and Problem Statement

A. Source Localization with overcomplete basis

For sensor array source localization problem, Our job is to find source locations given the information of the array geometry, the wave propagation medium parameters, and received signal from sensor array. we formalize the source localization with overcomplete representation problem following the notation in [19].

Consider K signal $u_k(t)$, $k \in \{1, ..., K\}$ as sources arrive at the omnidirectional sensor array M with additive noise $n_m(t)$, m-th sensor receives signal $y_m(t), m \in \{1, ..., M\}$ within the sensor array. The received signal can be expressed in (1), where $u(t)$ and $n(t)$ are defined in the similar fashion as $y(t)$.

$$y(t) = A(\theta)u(t) + n(t), t \in \{t_1, ..., t_T\} \tag{1}$$

The array manifold matrix $A(\theta)$ contains the delay and gain information for each source-sensor pair. Column vector $a(\theta_k)$, for $k \in \{1, ..., K\}$ is the steering vector. we are able to create the mapping $\theta \rightarrow A(\theta)$ by the array geometry and propagation parameters where $\theta = [\theta_1, ..., \theta_K]$. Given the received signal $y(t)$ and the mapping ability, the goal is to find the number of sources K and θ_k for all k. Estimating θ is not a simple linear estimation since the number of sources and the arrival angles are both unknown. The problem formulation above is based on the single snapshot for the sake of exposition simplicity. In fact, estimation based on joint-time sample can provide robustness to noise. we will formulate the source localization with overcomplete basis under the joint-time sample scenario.

One central assumption that most nonparametric methods rely on is signal sources can be treated as point sources and the number of sources are relatively small. This assumption naturally brings the overcomplete representation and sparse representation on the table. Note $\{\tilde{\theta_1}, ..., \tilde{\theta_{N_\theta}}\}$ as a sample grid of possible source locations, where N_θ is much larger than the number of sources K and number of sensors M. Then the array manifold matrix $A(\theta)$ can be rewrite as $A = [a(\tilde{\theta_1}), ..., a(\tilde{\theta_{N_\theta}})]$. Then $u(t)$ will need to change to a sparse spectrum $S = [s(t_1), ..., s(t_T)]$, similar to $n(t)$ and $y(t)$. The localization problem is reformulated to (2). And our goal changes from The nonlinear estimation of $A(\theta)$ to the estimation of S.

$$Y = AS + N \tag{2}$$

It is clear that (2) is ill-posed and has infinite solutions. Enforcing spatial sparsity of S can solve this problem. Solving (2) with l_0-norm that counts the number nonzero element is NP-hard. Instead, l_1-norm is proved to provide the exact solution when S is "sparse enough" with respect to A. And it can be solved by linear programming. Under the joint-time sample scenario, signals are not necessarily sparse in time. So we first calculate the l_2-norm of all time samples for each spatial index of S shown in (3), then enforce spatial sparsity using l_1-norm. The cost function can be described as in (4). For the real data, (4) can be solved with quadratic programming. And for the complex data, second order cone (SOC) programming can help to solve.

$$s_i^{(l_2)} = \|[s_i(t_1), ..., s_i(t_T)]\|_2 \tag{3}$$

$$\min \|Y - AS\|_2^2 + \lambda \|s^{(l_2)}\|_1 \tag{4}$$

Although formulating source localization problem with overcomplete basis and sparse representation is able to provide super resolution performance, robustness to noise, ability to solve coherent source problem, the main drawback is the computational complexity. In the

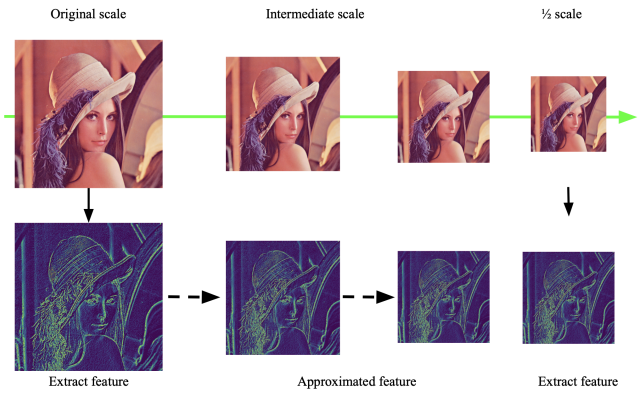

Fig. 1. Illustration of shift invariant feature approximation within an octave. Top row: image pyramid. Bottom row: feature pyramid. Solid and dotted arrows meas exact and approximated computation.

next section, we will introduce l_1-SVD to accelerate the computation.

B. Multiscale Feature Approximation

Multiscale feature pyramid provides robust feature representation and serves as the fundamental structure in many object detection algorithms. But computing exact feature pyramid is computational intense. Since most of the image structure is preserved in the resampled image, it is desirable to explore whether we can use one scale feature map to approximate the feature maps in the nearby scales illustrated in Fig. 1. How to construct the accurate approximation with a function of scale difference follows. Note Ω as a shift invariant function that generate a per-pixel feature map C_s given the $h_s \times w_s$ input image I_s at scale s. Widely used feature like gradient histogram, linear filter, color statistics etc. can be written as the weighted sum of C_s, the extracted feature shown in (5)

$$f_\Omega(I_s) \equiv \frac{1}{h_s w_s k} \sum_{ijk} C_s(i,j,k) \text{ , where } C_s = \Omega(I_s) \tag{5}$$

From previous work on image statistics, the expectation $E[\cdot]$ over ensemble images statistics $\phi(I)$ shows power relation across different scales [21], [22] shown in (6), where s_1, s_2 represent two different sclaes and λ_ϕ is an unique constant for each statistic.

$$E[\phi(I_{s_1})]/E[\phi(I_{s_2})] = (s_1/s_2)^{-\lambda_\phi} \tag{6}$$

$$f_\Omega(I_{s_1})/f_\Omega(I_{s_2}) = (s_1/s_2)^{-\lambda_\Omega} + \epsilon \tag{7}$$

For single image with shift invariant function, decomposing the image into K patches can approximate the expectation of $f_\Omega(I) \approx \sum f_\Omega(I_k)/K$, where $f_\Omega(I) \approx E[f_\Omega(I_k)]$. Eq. (6) can be rewritten as (7), where ϵ is the deviation from the power law for a given image. Now the task is to estimate λ_Ω. Compute the average scaling effect μ_s over the image samples(at scale s), $\mu_s = s^{-\lambda_\Omega} + E[\epsilon]$ can be derived from (7). Using least square fit on the log scale can easily estimate λ_Ω. Empirically results on several common shift invariant features (histogram of gradients, HOG etc.) show the expected error term is really small, and the deviation for individual images is also relatively small when scale change is small, which indicates the feature pyramid estimation is approachable. We will illustrate the proposed method in next section.

III. Methods

Although overcomplete representation brings a lot of advantages, an implicit disadvantage is the increase computation complexity. In the following subsections, we will illustrate the proposed approaches in both areas to decrease the computational complexity by approximating the feature representation.

A. l_1-SVD

To make the source localization with sparse representation tractable in the real world application, decomposing the joint-time sample into signal subspace and noise subspace can reduce the searching dimension introduced by joint time samples. Recall the received joint-time signal $Y = [y(t_1), ..., y(t_T)]$ is a $M \times T$ matrix. Y can be decomposed into K dimensional signal subspace and $T - K$ dimensional noise subspace by applying SVD: $Y = ULV'$. And we only keep the top-K dimensions that represent the signal subspace $Y_{SV} = YVD_k$, where $D_K = [I_K 0']$ that chooses the top K basis. Similarly, let $S_{SV} = SVD_K$ and $N_{SV} = NVD_K$. For each signal subspace singular vector, we have (8). K is usually much smaller than T, which brings significant computation complexity reduction. And we still impose l_1-norm to the spatial index of S_{SV} to ensure sparsity and apply l_2-norm across the singular index k, the objective function becomes (9), where λ can be chosen following discrepancy principle.

$$y^{SV}(k) = As^{SV}(k) + n^{SV}(k), k = 1, ..., K \tag{8}$$

$$\min \|Y_{SV} - AS_{SV}\|_2^2 + \lambda \|\tilde{s}^{(l_2)}\|_1 \tag{9}$$

Notice that $\|\tilde{s}^{(l_2)}\|_1 = \Sigma_{i=1}^{N_\theta}\sqrt{\Sigma_{k=1}^{K}(s_i^{SV}(k))^2}$ is neither linear or quadratic. SOC programming can solve the problem with complexity $O((K \times N_\theta)^3)$, without l_1-SVD, the complexity is $O((T \times N_\theta)^3)$, where $T \gg K$.

B. Fast Feature Pyramid

Feature pyramid ensembles multi-scale featuremaps as an overcomplete feature representation of the input image. Scales are evenly sampled in the log-space, with typically 4-12 scales per octave (an octave is the interval between one scale and its double). A standard way of calculating feature pyramid is to extract feature at every scale. From (7), the feature map C_s at scale s can be approximated by multiplying the nearest resized feature map $C_{s'}$ at scale s' and the estimated scaling factor $(s/s')^{-\lambda_\Omega}$: $C_s \approx R(C_{s'}, s/s')(s/s')^{-\lambda_\Omega}$, where $s' \in \{1, \frac{1}{2}, \frac{1}{4}, ...\}$ is the calculated nearest scale. To build the fast feature pyramid, feature map is calculated only once per octave. And the rest of the feature can be extrapolated with the corresponding nearest feature map.

Usually, calculating Ω is linear in number of pixels $n \times n$ (for simplicity), the computational complexity of building a feature pyramid with m scales per octave is $\sum_{k=0}^{\inf} n^2 2^{-2k/m} = \frac{n^2}{1-4^{-1/m}} \approx \frac{mn^2}{ln4}$. Typically, m is between 8 to 12. The fast feature pyramid approach only computes feature map once per octave, which achieved an order of magnitude reduction in computational complexity.

Both methods try to accelerate calculation from different angles: dimension reduction and feature extrapolation. In the next section, we will illustrate the achieved performance in both accuracy aspect and speed aspect.

IV. Evaluation

A. Source localization

To illustrate the performance superiority of l_1-SVD, the sensor array is set to be a $M = 8$ uniform linear array separated by half a wavelength, two zero-mean source signals in the far-field impinge on the array. Let $T = 200$, the grid is 1° sampled resulting $N_\theta = 180$.

In the first comparison shown in Fig. 2, two sources are closely located at 62° and 67°, which is within the Raleigh's limit. The experiment is set in a relatively noisy environment that $SNR = 0dB$. From Fig. 2, beamforming, Capon and MUSIC all merge two sources into one peak, except l_1-SVD. This comparison illustrates the super resolution ability and the robustness of noise.

Since l_1-SVD doesn't rely on the orthogonality assumption of signal and noise subspace, l_1-SVD is able

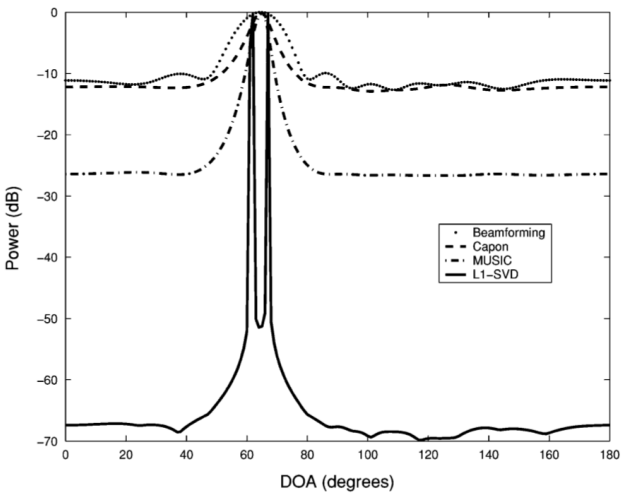

Fig. 2. Spatial Spectra for beamforming, Capon's, MUSIC and l_1-SVD for uncorrelated sources from [19]. DOAs: 62° and 67°. $SNR = 0dB$.

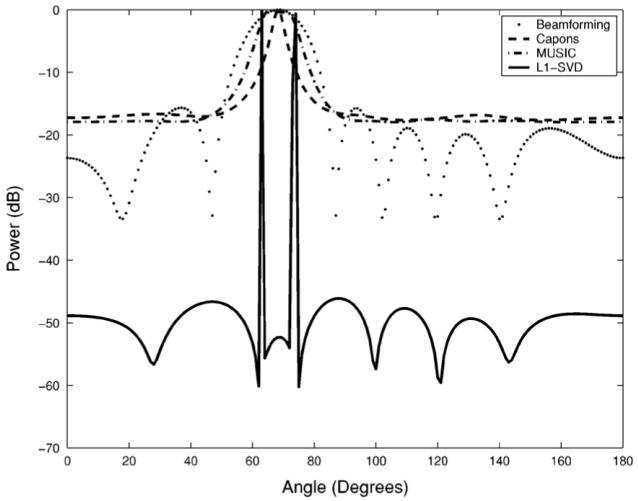

Fig. 3. Spectra for correlated sources from [19]. DOAs: 63° and 73°. $SNR = 20dB$.

to solve correlated signal scenarios shown in Fig. 3. Let two sources with correlation coefficient of 0.99 located at 63° and 73° at $SNR = 20dB$ environment. Again, only l_1-SVD is able to locate two sources, which prove its ability to resolve correlated sources case.

B. Fast feature Pyramid Application: Object Detection

The effectiveness of fast feature pyramid method compared to the traditional should be measured based two factors: accuracy and inference speed. Given the fact that the fast feature pyramid is nothing but a fast

TABLE I
MRs OF LEADING APPROACHES ON FOUR PEDESTRIAN DETECTION DATA SETS (INHERITED FROM [3])

	INRIA [2]	Caltech [6]	TUD	ETH [8]
VJ [24]	72	95	95	90
HOG [2]	46	68	78	64
Crosstalk [4]	19	54	58	52
ICF-exact	18	48	53	**50**
ICF	19	51	55	56
ACF-exact	**17**	**43**	**50**	**50**
ACF	**17**	45	52	51

version of traditional feature pyramid, we hope the object detection accuracy (log-average miss rate (MR)) should achieve the similar performance compared to normal feature pyramid. As for speed, since different object detection algorithm speed bottleneck varies, there is no unified measurement for the speed improvement. But the computation of exact feature pyramid runs at $\sim 15fps$, whereas the fast feature pyramid speeds up to nearly $50fps$ under the experiment system settings. From the computational complexity derived from the previous section, noticeable speed improvement is anticipated.

Three models are tested with the fast feature pyramid implementation:

- Aggregated Channel Features (ACF) aggregated normalized gradient magnitude, histogram of oriented gradients and LUV color space as the single pixel lookups as the extracted feature. Multiscale sliding window approach with boosted trees are used to detect objects within the frame.
- Integral Channel Features (ICF) as a precursor to ACF shares the same channel features and boosted tree strucutre. The key difference is ICF sums over rectangular channel regions instead of using aggregated pixel lookups.
- Deformable Part Models (DPM) [10] uses a variant of HOG features as the image representation, and a linear SVM as the classifer. The object model contains a coarse root model and optional finer components.

From Table. I, ACF-exact (normal feature pyramid) achieved best performance in pedestrian detection, and there is no significant performance drop with fast feature pyramid substitution. The direct comparison of DPM in pedestrian detection data sets is unavailable. But in general object detection dataset (VOC), the mean Average Precision (AP) is achieved 26.6 by exact feature pyramid and 24.5 by fast feature pyramid across 20 classes.

V. CONCLUSION AND DISCUSSION

Overcomplete representation brings huge performance improvement in computer vision and source localization area, the implicit computational complexity yet impede the tractability in the real world application. In this paper, we illustrate the problem statement with overcomplete representation in both vision area and source localization. With the complexity burden in mind, we describe two approximation methods- l_1-SVD and fast feature pyramid for acceleration. And we also evaluate their performance based on the traditional evaluation metrics in each area. For source localization, using overcomplete representation not only achieves the localization super resolution under noisy environment, but also solves the correlated sources scenario. In vision domain, feature pyramid provides robust features under various conditions, which already become a default feature representation structure. These approximation methods indeed reduce the complexity at least an order of magnitude, which make these approaches tractable.

The overcomplete representation is easy to introduce sparsity, which suits in the basic assumption of source localization there are only small number of sources present in the scene. We tested the performance of l_1-SVD using a powerful toolbox [25], the localization performance highly depends on the choice of l_1 regularization parameter λ. And the estimation of λ often does not yield the best results. Investigation of alternative way of utilizing sparsity is needed.

Feature pyramid provides robust features that brings huge performance improvement in detecting objects in various scales. But the DPM performance decrease in the general object detection is quite significant, additional experiment to examine the importance of different scales in detecting objects may better verify the validity of fast feature pyramid. Moreover, INRIA dataset that are used to empirically prove the validity of fast feature pyramid is not challenging enough compared to general detection dataset. Whether the power law of scale still holds in more complex scenes is also worthwhile to investigate. Also MR is not a complete measurement of detection performance, intersection over union (IOU) or ROC curve might be more insightful for presenting the performance variations.

REFERENCES

[1] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition*, volume 1, pages 886–893. Ieee, 2005.

[3] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545, 2014.

[4] P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *European conference on computer vision*, pages 645–659. Springer, 2012.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. *The British Machine Vision Conference*, 2009.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[7] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

[8] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[9] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2241–2248. Ieee, 2010.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] D. H. Johnson and D. E. Dudgeon. *Array signal processing: concepts and techniques*. Simon & Schuster, Inc., 1992.

[13] H. Krim and M. Viberg. Two decades of array signal processing research: the parametric approach. *IEEE signal processing magazine*, 13(4):67–94, 1996.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[19] D. Malioutov, M. Cetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE transactions on signal processing*, 53(8):3010–3022, 2005.

[20] M. Pedersoli, A. Vedaldi, J. Gonzalez, and X. Roca. A coarse-to-fine approach for fast deformable object detection. *Pattern Recognition*, 48(5):1844–1853, 2015.

[21] D. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

[22] D. L. Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4):517, 1994.

[23] R. O. Schmidt. *A signal subspace approach to multiple emitter location and spectral estimation*. Stanford University, 1982.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*, volume 1, pages I–I. Ieee, 2001.

[25] M. Wang, Z. Zhang, and A. Nehorai. Grid-less doa estimation using sparse linear arrays based on wasserstein distance. *IEEE Signal Processing Letters*, 26(6):838–842, 2019.