Detecting Source Video Artifacts with Supervised Sparse Filters

Todd R. Goodall and Alan C. Bovik, *Fellow*, *IEEE*

***Abstract*—A variety of powerful picture quality predictors are available that rely on neuro-statistical models of distortion perception. We extend these principles to video source inspection, by coupling spatial divisive normalization with a filterbank tuned for artifact detection, implemented in an augmented sparse functional form. We call this method the Video Impairment Detection by SParse Error CapTure (VID-SPECT). We configure VID-SPECT to create state-of-the-art detectors of two kinds of commonly encountered source video artifacts: upscaling and combing.**

The system detects upscaling, identifies upscaling type, and predicts the native video resolution. It also detects combing artifacts arising from interlacing. Our approach is simple, highly generalizable, and yields better accuracy than competing methods. A software release of VID-SPECT is available online: http://live.ece.utexas.edu/research/quality/VIDSPECT_release.zip for public use and evaluation.

***Index Terms*—VID-SPECT; Natural Scene Statistics; Upscaling prediction; Combing prediction; Interlace prediction; Sparsity; Sparse Filterbanks; Source Inspection**

I. Introduction

Source inspection is important for evaluating the quality of any large video collection. For this inspection task, a series of detectors may be used, where each detector is tuned to detect a specific artifact. A logical theoretical foundation is provided by natural scene statistics models, which are highly sensitive to picture distortions [1]. Such tools would be quite valuable to content providers such as Netflix, Hulu, and YouTube to assess their video collections, and to evaluate videos they ingest.

Sometimes video contents are upscaled during post-production, transcoding, or to fit larger formats. Upscaling artifacts are produced by imputing missing information from surrounding pixel data. This happens during color interpolation (demosaicking) and when adapting images for higher resolution displays. Since data imputation does not add information, and usually involves interpolation, upscaled images tend to be smoother than their originals, with reduced high-frequency energy. Upscaling a video results in lower dimensional data in a higher dimensional space.

Combing occurs when videos are represented in an interlaced form, where whole video frames are sequenced as "top-bottom" or "even-odd" frame pairs. Since the even-odd frame pairs are slightly temporally displaced in time, when they are reconstituted into whole frames, combing artifacts occur, particularly in regions of motion.

Upscaling prediction algorithms exist for (a) finding evidence of upscaling, (b) predicting native resolution, (c) classifying upscaling by type, and (d) quantifying perceptible loss of quality. Most existing methods do not fully cover this problem space, instead being designed to solve (a) or (b).

For (a), typical approaches involve covariance or radon transform analysis [2]. Periodicities introduced by upscaling have been deeply studied [3]–[8]. For (c), frequency-based approaches derive closed form predictions, but more general energy falloff-based models aided by machine learning better characterize differences amongst upscaling techniques [9] [10]. Methods that rely upon the Discrete Fourier Transform (DFT) typically lose prediction power when handling upscaling ratios outside the range of 1x-2x [10] [11].

Combing detectors have utilized top-field-first (TFF) and bottom-field-first (BFF) information across frames. For example, the interlace detector in FFmpeg [12] determines where the ratio TFF/BFF exceeds a threshold. Baylon [13] introduced a "zipper filter" to detect differences between TFF and BFF by analyzing moving edges. Each of these models requires more than one frame to affect detection, although the combing artifact is present in a single frame. Similar detectors are provided in [14]–[16].

We address several subproblems of upscaling and combing detection by learning sets of filters from pre-processed images using an augmented sparse functional. Our general model, called the Video Impairment Detector by SParse Error CapTure (VID-SPECT), computes predictions using averaged responses of filter-based feature extractors. We show that its detection performance significantly exceeds that of competitive models.

Section II describes a preprocessing model related to natural scene statistics and introduces the concept of developing sparse features tuned for artifact detection. Section III describes the upscaling artifact along with detection, classification, and native resolution prediction methods. Section IV describes the combing artifact detection method. Lastly, Section V presents concluding remarks.

II. Models

A. Natural Scene Statistic Pre-Processing Model

A variety of successful Image Quality Assessment (IQA) models utilize a normalizing transform, expressed as a local bandpass filtering operation followed by a local non-linear divisive normalization [17]. One such transform, known as the Mean-Subtracted Contrast Normalized (MSCN) transform, strongly Gaussianizes and decorrelates good quality photographic images [18]–[21]. This transform locally normalizes

spatial energy, similar to models of the retinal output signal, i.e. the contrast signal. The MSCN transform is usually defined

$$\hat{I}(\mathbf{x}) = \frac{I(\mathbf{x}) - \mu(\mathbf{x})}{\sigma(\mathbf{x}) + C}$$

where

$$\mu(\mathbf{x}) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(\mathbf{x})$$

and

$$\sigma(\mathbf{x}) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} (I_{k,l}(\mathbf{x}) - \mu(\mathbf{x}))^2},$$

where $K = L = 3$, $\mathbf{x}$ is the pixel location vector, and $w = \{w_{k,l} | k = -K, \cdots, K, l = -L, \cdots, L\}$ is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and normalized to unit volume. The parameter is commonly fixed at $C = 1$ to avoid saturating in low-contrast regions.

B. Sparse Functional Modeling

The entropy-reducing processing performed in mammalian visual cortex, which is fed by the retinal signal, has been hypothesized to have adapted optimally to efficiently encode images of natural scenes [22]. We are interested in developing similar optimal encoding schemes for specific visual detection tasks. Just as visual cortex can be modeled as an over-complete filterbank, we consider the possibility of learning embedded patterns in MSCN transformed images using an automatic feature extraction technique that utilizes such a learned filterbank. Towards this purpose, sparse dictionary learning can be used to discover those atoms which underlie pristine and distorted natural images.

Sparsity applied on image patches has shown utility in general recognition and denoising problems. The patch-based sparsity functional, which seeks to minimize the difference between batch of MSCN-trasformed patches S and a small number of weighted basis functions ϕ is defined by

$$\underset{X,\phi}{\operatorname{argmin}} \frac{1}{2} \left\| S - \sum_k \phi_k X_k \right\|_2^2 + \lambda \|X\|_1 \qquad (1)$$

subject to

$$\|\phi_k\|_2 = 1,\ X \geq 0.$$

Note that each basis function in ϕ is constrained to share the same dimension as the input MSCN patch S_i. Sparsity is achieved by penalizing the absolute sum of coding matrix X using an Lagrangian multiplier λ. This type of penalization of the coding matrix is known as the ℓ_1 norm.

Since this functional is unsupervised, it does not fully exploit additional information (such as labels). To overcome this, binary labels that indicate artifact presence may be added to the functional. The updated functional with labels is given by

$$\underset{X,\phi,p_c}{\operatorname{argmin}} \left[\frac{1}{2} \left\| S - \sum_k \phi_k X_k \right\|_2^2 - \alpha \sum_c y_c \log(p_c) + \lambda \|X\|_1 \right], \qquad (2)$$

subject to

$$\|\phi_k\|_2 = 1,\ X \geq 0$$

where the first term penalizes the reconstruction error and the second penalizes non-discriminative codes, y is a matrix of binary class labels, and $\|X\|_1$ is the sparsity term. The codes in X are constrained non-negative to enforce an additive relationship among unit-normalized dictionary elements. The predicted class label vector p_i is computed using a linear projection followed by softmax normalization, using

$$p_c = \frac{e^{\sum |SW_c \phi_c| + b_c}}{\sum_j e^{\sum |SW_j \phi_j| + b_j}}$$

to project correlations between filters and the input signal onto probability estimates. The diagonal weight matrix W_c is constrained nonnegative to enforce correlation between signal and ϕ while reweighing the contributions of each correlation to the overall prediction of class c. Finally, b is the class bias. The term $SW_c\phi$ measures correlation of reweighted dictionary elements $W_c\phi$ with the data S. The absolute value of this correlation is analogous to an activation function in neural networks and is maximized when ϕ correlates with the data in the assigned class. Values in W_c can be set to 0 to disable elements in ϕ for a class.

This approach to incorporating labels into the sparse functional is closest to the work of Mairal *et. al.* [23], but unlike Mairal *et. al.*, there is no direct dependence between the sparse coding problem and the classification problem. As a result, the dictionary learned by minimizing Equation 2 will recover the same codes found by minimizing Equation 1 with the same dictionary. We find that enforcing independence between the code update step and the dictionary update step is necessary for the artifact detection task.

C. VID-SPECT Model

In order to expand from patch-based to whole-frame analysis of artifacts, we consider the sparse filterbanks learned by appropriate minimization of equations 1 and 2 to be tuned for detecting artifacts and predicting artifact intensity. We developed the VID-SPECT model which uses this filterbank as a set of feature extractors. The processing stages of VID-SPECT are: computing the MSCN transform on the input image, using a precomputed filterbank designed by appropriately minimizing equation 2, convolving the MSCN transformed image by that filterbank, averaging responses, then mapping those averages to class labels, as depicted in Fig. 1. As we show, VID-SPECT performs well across multiple tasks.

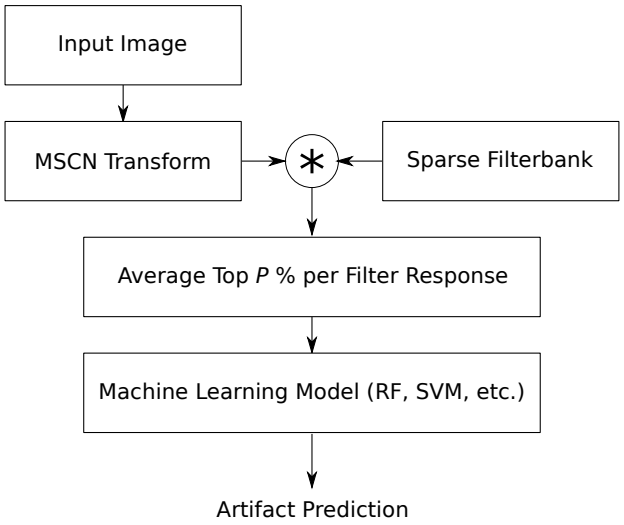


Fig. 1. Processing stages of VID-SPECT used to compute an artifact prediction given any input image and trained sparse filterbank.

III. Upscaling Problems

To study how natural video frames are perturbed by upscaling, we learned a set of sparse discriminative filters using both upscaled and non-upscaled video frames.

We studied four common upscaling interpolation schemes: bilinear, bicubic, Lanczos, and nearest neighbor upscaling. We collected a large dataset of more than 100,000 high quality Netflix video frames. We upscaled each frame using one of the interpolation functions. The upscaling ratios were randomly applied in the range [1.25, 3.0], since we wanted to include the practical extreme case where 720p film is upscaled to 2160p.

To generalize our upscaling analysis, we mixed two philosophies. First, we center cropped from within each video frame, then upscaled to the size of the frame, ensuring that pristine frame data was only perturbed by the upscaling artifact. In the alternative approach, we first downscaled the frame using a Lanczos-4 filter so that upscaling would maintain the original frame size, ensuring the content is held fixed across upscaling factors. We also consider frames downscaled using Lanczos-4 as a part of our non-upscaled frame data. These two scenarios were selected to alleviate concerns regarding scale in film content while also attempting to maintain upscaled film grain noise artifacts.

We then transformed frames using MSCN. For developing ϕ, we extracted several 25x25 patches from each frame, and used a Gaussian weighting function to suppress patch edges. This size of 25x25 was determined based on the maximum interpolation kernel width, which happens to be Lanczos kernel with upscaling factor of 3. This size also constrains the size of each element in ϕ. When evaluating VID-SPECT, we used patches of size 100x100. A total of 100,000 patches were used for training, and 60,000 patches were used for testing.

To explore the temporal aspect of videos, separate datasets for frame-differences were created using the same methodology. Two consecutive video frames are differenced then processed using MSCN before extracting patches. In this way, we directly compare the difference in prediction performance between single-frame and frame-difference detectors.

Towards understanding how well the learned system can characterize upscaling artifacts, we devised three tasks involving only the use of VID-SPECT features. The first task was to discriminate between upscaled and non-upscaled frames. The second involved identifying the interpolation scheme used from among non-upscaled (pristine), bilinear, bicubic, Lanczos, and nearest neighbor upscaling. The last task was to predict native resolution of both pristine and upscaled images.

A. Detection

Optimal parameters for the machine learning model in VID-SPECT are chosen by maximizing the median performance of 5-fold cross validation using just the training subset. To assess the binary classification performance, we measured the F1 score and the Matthews Correlation Coefficient (MCC). To assess regression performance, we measured Spearman's Rank-Ordered correlation Coefficient (SRCC) for monotonicity and Mean-Squared Error (MSE) for point-wise accuracy.

We evaluated VIDSPECT by choosing parameters for each model that reasonably spanned the parameter space for α and λ. We considered $\alpha \in \{0.0, 0.1, 1.0, 10.0\}$ and $\lambda \in \{0.1, 0.5, 1.0\}$. In each task, we constrained the number of filters in ϕ to be 100. Table I lists the performance results of VID-SPECT on the upscaling detection task. We tested both detection performance when only one interpolation method was present in the upscaled class, and also when all interpolation methods were present in the upscaled class. We also tested a version of VID-SPECT that uses frame differences rather than single frames, which we call VID-SPECT-D. From these results, we conclude that supervised single-frame VID-SPECT yielded the best upscaling detector.

B. Method Discrimination

The filters for the discrimination problem are provided in Fig. 2. These basis functions all exhibit directional high frequency patterns, which intuitively follows since upscaling artifacts mostly affect high-frequencies. The progression in interpolation order can be clearly seen across Bilinear, Bicubic, and Lanczos basis classes. In other words, bilinear basis functions exhibit patterns with 1-2 cycles, bicubic basis functions exhibit 2-3 cycles, and Lanczos exhibits at least two cycles, all at different orientations. Nearest neighbor filters exhibit high-frequencies along the cardinal directions.

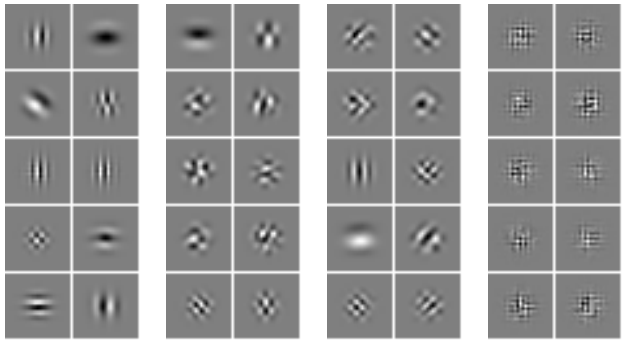
Table II compares performance across methods for the upscaling type discrimination task. Supervised VID-SPECT performed best compared against other models.

C. Native Resolution Prediction

Table III lists native resolution prediction performances across algorithms. Single-frame inputs to VID-SPECT delivered much better predictions of native resolution than the other models.

TABLE I
UPSCALING DETECTION PERFORMANCE ON VIDEO FRAME PATCHES WHERE UPSCALING TYPE INCLUDES "NOT UPSCALED," "BILINEAR UPSCALING," "BICUBIC UPSCALING," "LANCZOS UPSCALING," AND "NEAREST NEIGHBOR UPSCALING."

Algorithm	Bilinear		Bicubic		Lanczos		Nearest Neighbor		All	
	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC
VID-SPECT ($\alpha = 1.0$)	**0.9950**	**0.9899**	**0.9949**	**0.9897**	**0.9952**	**0.9904**	0.9923	0.9845	**0.9909**	**0.9818**
VID-SPECT ($\alpha = 0.0$)	0.9715	0.9427	0.9843	0.9686	0.9931	0.9862	0.9810	0.9620	0.9689	0.9379
VID-SPECT-D ($\alpha = 10.0$)	0.9884	0.9767	0.9909	0.9819	0.9934	0.9868	0.9914	0.9827	0.9875	0.9750
VID-SPECT-D ($\alpha = 0.0$)	0.9860	0.9719	0.9894	0.9788	0.9926	0.9853	0.9884	0.9768	0.9847	0.9693
Goodall *et al.* [24]	0.9872	0.9744	0.9885	0.9769	0.9941	0.9882	**0.9977**	**0.9953**	0.9893	0.9786
BRISQUE [19]	0.9331	0.8650	0.8988	0.7949	0.8847	0.7657	0.8847	0.7639	0.8730	0.7417
Vázquez-Padín *et al.* [25]	0.9736	0.9469	0.9706	0.9409	0.9683	0.9361	0.9929	0.9858	0.9729	0.9454
Feng *et al.* [10]	0.8609	0.7207	0.9162	0.8303	0.9577	0.9155	0.9099	0.8150	0.8555	0.7206



(a) Bilinear (b) Bicubic (c) Lanczos (d) Neighbor

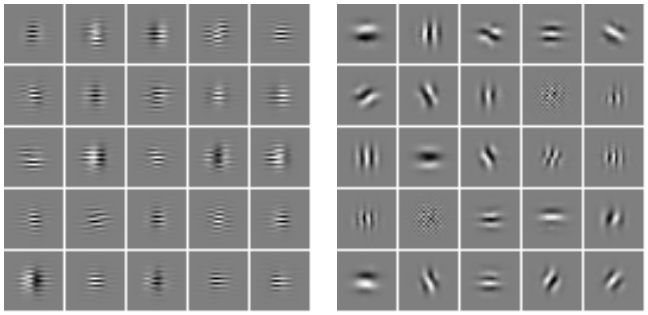
Fig. 2. Dictionaries learned for each evidence category, when assigning 10 filters to each. Filter size is held constant at 25x25.

TABLE II
UPSCALING TYPE DISCRIMINATION PERFORMANCE ON VIDEO FRAME PATCHES WHEN CLASSIFYING UPSCALING TYPE AMONG "NOT UPSCALED," "BILINEAR UPSCALING," "BICUBIC UPSCALING," "LANCZOS UPSCALING," AND "NEAREST NEIGHBOR UPSCALING." REPORTED VALUES ARE F1-MACRO SCORES, SINCE THE CLASSES ARE WELL-BALANCED.

Algorithm	F1-Macro
VID-SPECT ($\alpha = 1.0$)	**0.9225**
VID-SPECT ($\alpha = 0.0$)	0.9206
VID-SPECT-D ($\alpha = 10.0$)	0.8965
VID-SPECT-D ($\alpha = 0.0$)	0.8838
Goodall *et al.* [24]	0.8753
BRISQUE [19]	0.4921
Feng *et al.* [10]	0.7519

TABLE III
NATIVE RESOLUTION PREDICTION ON PATCHES THAT WERE NOT UPSCALED, AND UPSCALED USING "BILINEAR UPSCALING," "BICUBIC UPSCALING," "LANCZOS UPSCALING," AND "NEAREST NEIGHBOR UPSCALING" WITH UPSCALING RATIOS CHOSEN FROM THE RANGE 1.25X TO 3X.

Algorithm	MSE	SRCC
VID-SPECT ($\alpha = 0.1$)	**26.28**	**0.9445**
VID-SPECT ($\alpha = 0.0$)	27.20	0.9353
VID-SPECT-D ($\alpha = 10.0$)	37.85	0.9250
VID-SPECT-D ($\alpha = 0.0$)	46.13	0.9179
Goodall *et al.* [24]	70.70	0.9055
BRISQUE [19]	282.86	0.7663
Vázquez-Padín *et al.* [25]	227.66	0.8591
Feng *et al.* [10]	238.02	0.8048
Pfennig and Kirchner [11]	505.33	0.6184



(a) Positive evidence (b) Negative evidence

Fig. 3. Sparse filters learned for interlacing.

IV. INTERLACE DETECTION

Combing artifacts can be much more visually obvious than upscaling effects when viewed on progressive displays. The artifact often becomes increasingly obvious on scenes containing rapid motion.

We collected a training/validation dataset of 581 interlaced combed sequences of 3 frames. A combed sequence is one where the middle frame exhibits visible combing (the others may also). To balance these positive samples, an equally sized set of 581 non-interlaced video sequences was gathered as negative examples. A negative sequence is one where no frames exhibit visible combing. We collected a separate content-distinct test dataset containing 75 interlaced three-frame sequences and 75 undistorted three-frame sequences.

Figure 3 depicts evidence for and against the interlacing artifact. In Fig. 3a, the zigzag pattern of combing is apparent. In Fig. 3b, low-frequencies and vertical edges dominate.

We evaluated two existing state-of-the-art algorithms. The first is the FFmpeg 'idet' detector, which requires 3 frames. For progressive video, it assumes that the row in the current frame can be interpolated using two rows in either the previous or next frame. For interlaced video, it assumes the interpolated row will not match the corresponding row in the previous or next frames. A prediction is generated by applying threshold T_1 on these two measurements.

TABLE IV
COMBING DETECTION RESULTS COMPUTED ON THE TEST SET OF 150 VIDEO SEQUENCES.

Algorithm	F1	MCC
VID-SPECT ($\alpha = 1.0$)	**0.9730**	**0.9470**
VID-SPECT ($\alpha = 0.0$)	**0.9730**	**0.9470**
VID-SPECT-D ($\alpha = 10.0$)	0.9306	0.8695
VID-SPECT-D ($\alpha = 0.0$)	0.9241	0.8552
BRISQUE [19]	0.8718	0.7357
FFmpeg	0.9167	0.8427
Baylon [13]	0.8811	0.7761

The second algorithm was developed to determine field order on known combed sequences [13]. We modified it to provide detection predictions. It counts the number of zipper artifacts T_0 of length Z in the top-field and the bottom field between two frames. If the difference between these counts exceeds a threshold T_1, then the two frames are labeled as combed. Thus, this algorithm requires two frames for detection.

Table IV lists the F1 and MCC performances for the selected algorithms. The VID-SPECT detector yielded higher accuracy than the compared detectors. The BRISQUE quality model performed almost as well as Baylon's detector, despite not being designed for this artifact. We determined thresholds in the FFmpeg 'idet' detector and in Baylon's detector using 5-fold cross validation. The optimal threshold parameter for FFmpeg's detector was $T_1 = 1.0551$, and the optimal parameters for Baylon's detector were $T_0 = 75$, $T_1 = 1.113$, and $Z = 10$.

V. CONCLUSION/FUTURE WORK

We proposed a new, general-purpose video source inspection framework called VID-SPECT, which uses sparse basis functions computed on MSCN coefficients as feature extractors to detect two types of source artifacts, upscaling and combing. We recommend that VID-SPECT be configured to use basis functions derived from a supervised functional to find best discriminative filters. Given the effectiveness of VID-SPECT on upscaling and combing problems, video engineers should be able to extend this method to solve a wider array of artifact detection problems, from subtle artifacts to more obvious artifacts.

ACKNOWLEDGEMENTS

We thank Netflix for providing necessary access to data and research guidance. We also thank NVIDIA for providing a Tesla K40 GPGPU, which we used to accelerate convolution operations. Finally, we acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu.

REFERENCES

[1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[2] B. Mahdian and S. Saic, "Blind authentication using periodic properties of interpolation," *IEEE Trans. Info. Forensics Sec.*, vol. 3, no. 3, pp. 529–538, 2008.

[3] A. C. Gallagher, "Detection of linear and cubic interpolation in jpeg compressed images," *Canadian Conf. Computer Robot Vision*, pp. 65–72, 2005.

[4] S. Prasad and K. Ramakrishnan, "On resampling detection and its application to detect image tampering," *IEEE Int'l. Conf. Multimedia Expo*, pp. 1325–1328, 2006.

[5] S.-J. Ryu and H.-K. Lee, "Estimation of linear transformation by analyzing the periodicity of interpolation," *Pattern Recog. Lett.*, vol. 36, pp. 89–99, 2014.

[6] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, 2005.

[7] D. Vázquez-Padín and F. Pérez-González, "Prefilter design for forensic resampling estimation," *IEEE Int'l. Wkshp. Info. Forensics Sec.*, pp. 1–6, 2011.

[8] M. Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue," *ACM Wkshp. Multimedia Sec.*, pp. 11–20, 2008.

[9] I. Katsavounidis, A. Aaron, and D. Ronca, "Native resolution detection of video sequences," *Soc. Motion Picture Television Engrs.*, 2015.

[10] X. Feng, I. J. Cox, and G. Doerr, "Normalized energy density-based forensic detection of resampled images," *IEEE Trans Multimedia*, vol. 14, no. 3, pp. 536–545, 2012.

[11] S. Pfennig and M. Kirchner, "Spectral methods to determine the exact scaling factor of resampled digital images," *Int'l. Symp. Comm. Control Signal Process.*, pp. 1–6, 2012.

[12] "Interlace Detector (idet)," ffmpeg.org/ffmpeg-filters.html#idet, FFmpeg, accessed Mar 2017.

[13] D. M. Baylon, "On the detection of temporal field order in interlaced video data," *IEEE Int'l. Conf. Image Process.*, vol. 6, pp. VI–129, 2007.

[14] Y. Hui, "Progressive/interlace and redundant field detection for encoder," 2005, uS Patent 6,870,568. [Online]. Available: https://www.google.com/patents/US6870568

[15] M. Pindoria and T. Borer, "Automatic interlace or progressive video discrimination," *SMPTE Ann. Tech. Conference & Exhibition*, pp. 1–8, 2012.

[16] S. Keller, K. S. Pedersen, and F. Lauze, "Detecting interlaced or progressive source of video," in *Multimedia Signal Process.* IEEE, 2005, pp. 1–4.

[17] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

[18] D. L. Ruderman, "The statistics of natural images," *Network: Comput Neural Syst*, vol. 5, no. 4, pp. 517–548, 1994.

[19] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[20] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[21] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.

[22] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[23] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[24] T. Goodall, I. Katsavounidis, Z. Li, A. Aaron, and A. Bovik, "Blind picture upscaling ratio prediction," *IEEE Signal Process Lett*, vol. 23, no. 12, pp. 1801–1805, 2016.

[25] D. Vázquez-Padín, P. Comesaña, and F. Pérez-González, "An SVD approach to forensic image resampling detection," *EUSIPCO*, pp. 2067–2071, 2015.