Sequential Score Adaptation with Extreme Value Theory for Robust Railway Track Inspection

Xavier Gibert
University of Maryland
College Park, MD
gibert@umiacs.umd.edu

Vishal M. Patel
Rutgers University
Piscataway, NJ
vishal.m.patel@rutgers.edu

Rama Chellappa
University of Maryland
College Park, MD
rama@umiacs.umd.edu

Abstract

Periodic inspections are necessary to keep railroad tracks in state of good repair and prevent train accidents. Automatic track inspection using machine vision technology has become a very effective inspection tool. Because of its non-contact nature, this technology can be deployed on virtually any railway vehicle to continuously survey the tracks and send exception reports to track maintenance personnel. However, as appearance and imaging conditions vary, false alarm rates can dramatically change, making it difficult to select a good operating point. In this paper, we use extreme value theory (EVT) within a Bayesian framework to optimally adjust the sensitivity of anomaly detectors. We show that by approximating the lower tail of the probability density function (PDF) of the scores with an Exponential distribution (a special case of the Generalized Pareto distribution), and using the Gamma conjugate prior learned from the training data, it is possible to reduce the variability in false alarm rate and improve the overall performance. This method has shown an increase in the defect detection rate of rail fasteners in the presence of clutter (at PFA 0.1%) from 95.40% to 99.26% on the 85-mile Northeast Corridor (NEC) 2012-2013 concrete tie dataset.

1. Introduction

In sequential inspection problems, such as visual railway track inspection, a video feed is streamed from one or more cameras to a detection system, and we are interested in designing a detector that can find abnormal patterns in such data. There is a limit to the number of false alarms that the operator can handle, so it is necessary to select the optimal operating point at which the false alarm rate does not exceed such limit. Indeed, most of the data that an autonomous inspection vehicle will collect will be discarded without anyone ever looking at it. Therefore, an excessively high false alarm rate will result in a waste of storage space and bandwidth. The only relevant images are the ones that correspond to unexpected patterns, so we are actually interested in finding such anomalous patterns.

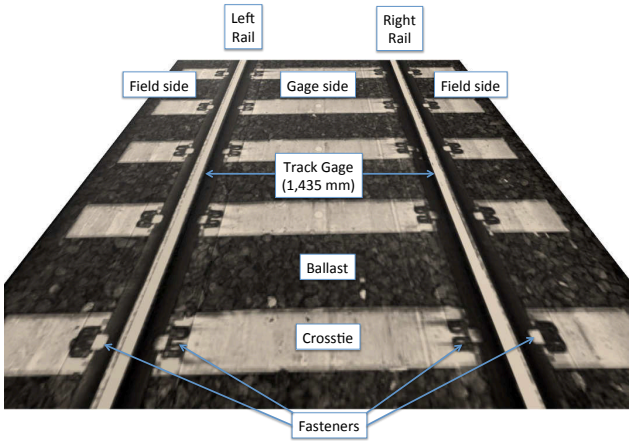


Figure 1. Definition of basic track elements.

Anomaly detection is a hypotheses testing problem in which the null hypothesis is that an image is normal and the alternative hypothesis is that it is anomalous. Due to the complexity of the scene and image formation process, both hypothesis are composite, with nuisance parameters arising from changes in illumination, occlusion, background clutter, and many other uncontrollable factors. Rather than trying to model each of these variables individually, in this paper we adapt the detection scores with the objective of reducing the variability in type I error rate. The is known as constant false alarm rate (CFAR) detection. We adopt the Bayesian view that such parameters are random variables with one realization per image. The images have a natural order based on the time they were captured at, so the sequence of these random parameters forms a random process. A key observation is that this random process has strong long-term dependencies. The effect of such slowly varying nuisance parameters is that false alarms are concentrated in small segments of the image sequence.

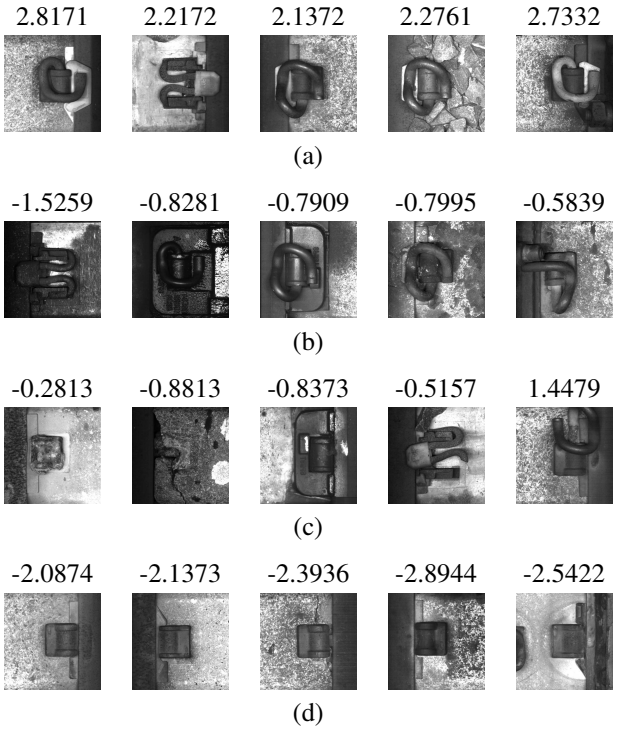


Figure 2. Examples of fastener scores (a) Good fasteners with high scores (b) Good fasteners with low scores (c) Defective fasteners with high scores (d) Defective fasteners with low scores

Figure 1 shows the definitions of several track components. In this paper, we focus on fastener inspection. Figure 2 shows examples of good and defective fasteners and their detection scores generated by the multi-task learning method [3] of Gibert *et al*. Although most fasteners have high scores and most defective ones have low scores, when good fasteners have low scores, there is an underlying phenomenon that causes scores of nearby images to also be low.

The rest of the paper is organized as follows. In Section 2 we review related works. The algorithm is described in Section 3. Experimental results are described in Section 4. Section 5 concludes the paper with a brief summary and discussion.

2. Background

2.1. Robust Anomaly Detection

The presence of outliers is a challenge that many computer vision systems have to deal with. The RANdom SAmple Consensus (RANSAC) algorithm [2] has been used in many applications for removing outliers when fitting a model to data. This method is specially useful when most of the samples follow a linear model plus additive i.i.d. Gaussian noise, but a few samples with gross errors do not follow this model. However, in many applications, it not clear which samples should be treated as inliers and which of them are outliers. For instance, in big data applications, the data just appears to have a distribution with long tails that decay at slower rate than the corresponding Gaussian distribution that best fits the data in the least squares sense. Indeed, what appears to be an outlier in feature space may just be a normal sample that has been subject to some kind of degradation for which the feature extractor was not designed for. These degradation modes may include impulse noise, partial occlusion, and in some cases, changes in appearance due to blur, shadows, or pose. In anomaly detection problems, the samples of interest are those in the tail of such data distribution. Therefore, any method that discards outliers have the potential of discarding anomalies, so in order to successfully find anomalies in such images it is necessary to use other methods.

The field of robust statistics [7, 10] provides the tools for estimation of unknown quantities when the underlying probability distribution is non-Gaussian and it is not known exactly. In practice, the data can be modeled as the mixture of a Gaussian distribution and a heavy-tailed distribution (the contaminated Gaussian model). In this case, it is desirable to design an estimator whose performance is minimax over a family of distributions that includes the Gaussian as a special case. There are basically three types of robust estimates: M-estimates[6] (Maximum likelihood type), L-estimates (Linear combination of order statistics), and R-estimates (Estimates derived from rank tests).

In supervised learning problems, there is a distinction on how to handle outliers at training time vs. testing time. Supervision at training time usually mitigates the problem of outliers as it is possible to manually select the inliers. The use of the ℓ_1 minimization promotes a sparse representation of the data. The solution of the ℓ_1 minimization is the Maximum Likelihood Estimate of the location parameter when the data follows a Laplacian distribution, and a straightforward way of robustifying a regression procedure is by replacing the ℓ_2 norm in the cost function by the ℓ_1 norm. A related L-estimator that results from such ℓ_1 optimization is the Least Median of Squares (LMS), which was introduced in the computer vision field by Kim *et al*. [8]. The drawback of the LMS is that the median estimator's efficiency is only $\frac{2}{\pi} = 0.637$ when the true distribution is Gaussian. The M-estimator based on the Huber loss function[6]

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & \text{for } |t| < k \\ k|t| - \frac{1}{2}k^2 & \text{for } |t| \geq k \end{cases} \quad (1)$$

is more flexible because it has the sample mean ($k = \infty$) and sample median ($k = 0$) as special cases and it can be tuned to handle different degrees of contamination in the contaminated Gaussian model. However, since this estimator depends on a scale parameter k (unlike L-estimators,

which are scale-invariant), it is necessary to first estimate this parameter using a robust scale estimator.

2.2. Extreme Value Theory for Adaptive Anomaly Detection

Due to illumination and viewpoint changes, clutter distribution, and other image degradation, the distribution of features extracted from images at test time, does not match what was observed during training. Moreover, such distribution may not be stationary, but slowly changes over time, so a fixed threshold would result in large variability in the false alarm rate. Broadwater and Chellappa[1] proposed a technique to find adaptive thresholds for Constant False Alarm Rate (CFAR) detectors based on Extreme Value Theory (EVT) [5] that can be used even when limited training data is available. EVT is applicable to problems where the probability of a rare event must be estimated even if such a rare event has never occurred. Scheirer *et al.* [12, 13] also used EVT for score normalization and showed its applicability to sensor fusion problems.

For completeness, we recall the EVT basic results below. Let $X_1, \ldots, X_n$ be i.i.d. samples from an unknown distribution F and $M_n = \max(X_1, \ldots, X_n)$, the maximum of n i.i.d. variables. The fundamental EVT theorem, the Fisher-Tippett-Gnedenko theorem[5], states that if there exist a sequence of pairs of real numbers (a_n, b_n) such that $a_n > 0$ for all n and a distribution function $\Lambda(x)$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \Lambda(x), \tag{2}$$

for all x at which $\Lambda(x)$ is continuous, then the limit distribution $\Lambda(x)$ belongs to either the Gumbel, the Fréchet or the Weibull family. These three families can be grouped into the Generalized Extreme Value Distribution (GEVD)

$$\Lambda(x; \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ the scale parameter and $\xi \in \mathbb{R}$ the shape parameter. The Gumbel distribution is a special case of the GEVD when $\xi = 0$, the Fréchet when $\xi > 0$, and the Weibull when $\xi < 0$. When the limiting distribution exists, we say that $F(x)$ lies in the "domain of attraction" of $\Lambda(x)$.

In many practical applications, we are interested in the tail distribution of the distribution F. Given an upper threshold u, we select the N_n samples that exceed such threshold and define the excesses $Y_1, \ldots, Y_{N_n}$ as $Y_i = X_j - n$, where i is the excess index and j is the index of the original sample. The probability of exceeding the threshold is $\lambda = 1 - F(u)$. For sufficiently large u, the upper tail distribution function $F_u(y)$ (the conditional distribution function of the excesses),

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} \tag{4}$$

can be approximated by a Generalized Pareto Distribution

$$G(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \quad y > 0. \tag{5}$$

where $\sigma > 0$, $\xi \in \mathbb{R}$, and $x_+ = \max(x, 0)$. This approximation is justified by the Pickands theorem[11], which states that

$$\inf_{\xi} \lim_{u\uparrow\omega_F} \inf_{\sigma} \sup_{y>0} |F_u(y) - G(y; \sigma, \xi)| = 0 \tag{6}$$

if and only if F is in the domain of attraction of the GEVD. Note that the exponential distribution is a special case of the GPD for $\xi = 0$, i.e. $G(y; \sigma, 0) = 1 - e^{-y/\sigma}$.

These results can be extended to the multivariate case, for example to model the tail distribution of the maximum of a cluster of observations. Under stationarity of observations, this can be achieved by incorporating both the tail of the marginal distribution and the so-called extremal index. Let $\{X_n : n \geq 1\}$ be a (strictly) stationary sequence of r.v.'s with marginal distribution F. Then, for sufficiently large n

$$P\{M_n \leq u_n\} \approx F^{n\theta}(u_n), \tag{7}$$

where u_n is any high threshold such that $n(1-F(u_n))$ converges to a positive number as $n \to \infty$ and θ is a fixed number in $[0, 1]$. θ is the extremal index that measures the strength of dependence of $\{X_n\}$. If $\{X_n\}$ are independent, then $\theta = 1$. On the other hand, if $\{X_n\}$ are highly dependent, then $\theta \approx 0$. A method for estimating the extremal index for a real-valued Markov chain was proposed by Yun [15].

3. Proposed Approach

In this section we describe our approach for normalizing the scores of an anomaly detector deployed in an application in which the distribution of the normal samples gradually changes over time. This may be caused by changes in illumination, change in view-point, addition or removal of clutter, or other uncontrollable factors. The approach is similar to the method proposed by Broadwater and Chellappa[1] in which an adaptive threshold is estimated from the GPD fit to the upper tail of the distribution after removing the outliers or targets using a Kolmogorov-Smirnov statistical test. The difference is that our method is Bayesian and we work with sequential data and estimate the adaptive threshold for each sample.

3.1. Bayesian Model

We want to adapt the scores of an anomaly detector applied to a sequence of images so that, when we apply a given threshold, we get an approximate CFAR. The images have been collected from a moving vehicle, so the environmental conditions and clutter distribution are not stationary, but slowly change over time. In EVT-based threshold estimation, it is necessary to estimate the parameters σ and ξ of the GPD from the upper- or lower-tail of the empirical distribution. For the rest of this paper we will refer to the upper tail of the distribution of the random variable X, but the same applies to the lower tail since the lower tail of X is the upper tail of $Z = -X$. The threshold u needs to be set high enough so that the tail of $F(x)$ converges in distribution to the GPD. However, since we are dealing with a non-stationary random process, we need to work on a small window centered at the sample of interest. This window needs to be long enough so that we can fit the parameters of the GPD to its tail (for example the largest 5% of the samples), but short enough that the distribution has not changed much. In applications in which the dynamics of the process change quickly, our options are rather limited. If we fit a GPD to the extreme samples of a short window, the estimated threshold has so much variance that the resulting performance is worse than using a fixed threshold. On the other hand, if the window is too long, the threshold does not adapt at all. For example, if we use a window of 100 samples and select the upper threshold to the 95th percentile, we would only have 5 samples to estimate the 2 parameters of the GPD, resulting in severe overfitting.

To overcome this limitation, we will make one simplification by fixing $\xi = 0$, so we only need to estimate one parameter instead of two. Under $\xi = 0$, the GPD reduces to the exponential distribution

$$G(y; \sigma, \xi = 0) = 1 - e^{-y/\sigma}. \tag{8}$$

For convenience, we apply the parameterization $\lambda = 1/\sigma$ and write the Exponential in its canonical form

$$G(y; \lambda) = 1 - e^{-\lambda y} \tag{9}$$

$$g(y; \lambda) = \lambda e^{-\lambda y}. \tag{10}$$

As opposed to the general case of the GPD, the Exponential distribution is a member of the exponential family, so it has a non-trivial sufficient statistic from which we can easily compute the maximum likelihood estimate (MLE) of its parameter. Its conjugate prior is the Gamma distribution,

$$\pi(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \tag{11}$$

the non-informative (improper) prior is given by $\alpha = 1$, $\beta = 0$, and the parameters of the Gamma posterior under a Gamma$(\lambda; \alpha_0, \beta_0)$ prior can be computed as

$$\alpha_1 = \alpha_0 + n \tag{12}$$

$$\beta_1 = \beta_0 + \sum_{i=1}^{n} y_i. \tag{13}$$

Moreover, the maximum a posteriori (MAP) estimate has the closed from $\widehat{\lambda} = \frac{\beta}{\alpha - 1}$. This simplified model allows us to derive a very fast adaptation algorithm that we describe in the following section. This approximation works well in practice, especially when the scores are trained with a sparsity promoting loss function such as the hinge loss.

Algorithm 1 EVT training algorithm.

procedure TRAIN($\mathcal{T}$, p_u, w_0)
  $n \leftarrow 0$, $s \leftarrow 0$ ▷ Initialize sufficient statistics
  for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}$ **do** ▷ Training set $\mathcal{T}$ contains $\mathbf{x}$ scores, $\mathbf{y}$ labels
    $\mathbf{g} \leftarrow \{x_i \mid y_i = 0\}$ ▷ Select negative samples
    $u \leftarrow u \mid \#\{g_i > u\} = \#\mathbf{g}\ p_u$ ▷ Find upper threshold
    $\mathbf{t} \leftarrow \{g_i \mid g_i > u\}$ - u ▷ Extract upper tail
    $n \leftarrow n + \#\mathbf{t}$ ▷ Update counts
    $s \leftarrow s + \sum \mathbf{t}$ ▷ Update sum
  end for
  $\alpha_0 \leftarrow 1 + w_0$
  $\beta_0 \leftarrow \frac{w_0\, s}{n}$
  return α_0, β_0 ▷ Parameters of the Gamma prior
end procedure

3.2. Training

Our training set $\mathcal{T}$ contains a number of sequences of scores $\mathbf{x}$ with their corresponding sequences of labels $\mathbf{y}$. During training, we compute the sufficient statistics n and s for all the samples that are not labeled as anomalies (the sufficient statistic is all we need to characterize the Gamma prior distribution). We then re-scale them to limit the effect of this prior. Effectively, we use w_0 pseudo-samples instead n (the number of samples in the training set). This is necessary because n is usually a very large number, and computing α_0 and β_0 with it would result in a very strong prior that would introduce too much bias in the MAP estimate.

The steps of the training procedure are described in Algorithm 1. The parameter p_u is the probability of the tail, and w_0 is the weight in sample counts that we assign to the training set. In our experiments we used $p_u = 0.05$ and $w_0 = 400$.

3.3. Proposed Adaptive Thresholding Algorithm

During testing, we first perform a series of Kolmogorov-Smirnov (KS) tests [14] to find and remove anomalies. The

Algorithm 2 EVT adaptive thresholding algorithm

procedure ADAPTSCORES($\mathbf{x}$, α_0, β_0, p_u, p_f, w_1, L, n_a)
  $\widehat{\lambda}_0 \leftarrow \frac{\beta_0}{\alpha_0 - 1}$ ▷ MLE in training set
  $\mathbf{y} \leftarrow$ **sort_desc**($\mathbf{x}$) ▷ Sort scores in descending order
  $k \leftarrow \#\mathbf{y}\ p_u$
  for $i \leftarrow 1, n_a$ **do** ▷ Training set $\mathcal{T}$ contains $\mathbf{x}$ scores, $\mathbf{y}$ labels
    $u \leftarrow y_{i+k}$ ▷ Find upper threshold
    $\mathbf{t} \leftarrow \{y_i, \ldots, y_{i+k}\} - u$ ▷ Extract upper tail
    $D_{n,i} = \sup_{x \in \mathbf{t}} \left| \widehat{G}_n(x) - G(x; \lambda) \right|$ ▷ Compute KS statistic
  end for
  $\hat{i} \leftarrow \min_i \{D_{n,i}\}$ ▷ Estimate number of outliers
  $u' \leftarrow y_{\hat{i}}$ ▷ Set outlier rejection threshold
  $\mathbf{t} \leftarrow \{y_{\hat{i}}, \ldots, y_{\hat{i}+k}\} - u$ ▷ Extract upper tail
  $\alpha_1 \leftarrow \alpha_0 + w_1$
  $\beta_1 \leftarrow \beta_0 + \frac{w_1 \sum \mathbf{t}}{\#\mathbf{t}}$
  for $i \leftarrow 1, n$ **do**
    $\mathbf{w} \leftarrow \mathbf{x}_{i-(L-1)/2:i+(L-1)/2}$ ▷ Window centered at sample x_i
    $u \leftarrow u \mid \#\{w_i > u\} = \#\mathbf{w}\ p_u$ ▷ Find upper threshold
    $\mathbf{t} \leftarrow \{w_i \mid w_i > u\}$ - u ▷ Extract upper tail
    $\alpha \leftarrow \alpha_1 + \#\mathbf{t}$ ▷ Posterior
    $\beta \leftarrow \beta_1 + \sum \mathbf{t}$ ▷ Posterior
    $\widehat{\lambda} \leftarrow \frac{\beta}{\alpha - 1}$ ▷ MAP estimate
    $y_i \leftarrow x_i + u - \widehat{\lambda}\ log(p_f / p_u)$ ▷ Adapt score
  end for
  return $\mathbf{y}$ ▷ Adapted scores
end procedure

KS statistic

$$D_n = \sup_x \left| \widehat{G}_n(x) - G(x; \widehat{\lambda}) \right| \tag{14}$$

measures the dissimilarity between distributions $G(x; \widehat{\lambda})$ and $\widehat{G}_n(x)$. $G(x; \widehat{\lambda})$ is the GPD in (9) and

$$\widehat{G}_n(x) = 1 - \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{15}$$

where $I(x)$ is a standard indicator function, is the empirical tail CDF. The KS test requires estimating a threshold K_α for rejecting (with confidence $1 - \alpha$) the hypothesis that the observed data does not fit G with the test $\sqrt{n} D_n > K_\alpha$. The result from Lilliefors[9] shows that the KS test is biased when the reference distribution G is not precisely known (in this case, $\widehat{\lambda}$ is estimated from the training data). However, as noted in [1], it is not necessary to identify the exact value of α for the purpose of removing anomalies and outliers. Instead, we first compute D_n with all the samples in the tail. We call this $D_{n,1}$. We then remove the largest sample and we compute $D_{n,2}$ using the remaining samples. We keep iterating until we get D_{n,n_a}. Finally, we select the value of i that minimizes $D_{n,i}$.

After removing the anomalies, we use the prior estimated during training to compute the posterior for the whole sequence. This posterior is used as the prior for estimating the tail distribution on each shift of a window centered on each of the samples. The details of the adaptation procedure are described in Algorithm 2. The input to the adaptation procedure is a sequence of scores $\mathbf{x}$, the parameters of the prior Gamma distribution α_0 and β_0, the size of the upper tail p_u, the target false alarm rate p_f, the weight w_1 that we assign to the the prior contribution of the whole sequence, the window length L, and the maximum number of anomalies n_a in the sequence. The output sequence $\mathbf{y}$ has been adapted so that when it is thresholded at 0, the false alarm rate is p_f. In our experiments, we have used $p_u = 0.05$, $p_f = 0.001$, $w_1 = 100$, $L = 101$, and $n_a = 12$.

4. Experimental Results

To validate the effectiveness of the proposed approach, we have used the 340 sequences of fastener detections corresponding to each of the 4 cameras in each of the 85 miles of the Amtrak NEC concrete tie dataset introduced in [4]. This dataset contains a total of 203,287 ties and each tie is divided in 4 regions (left field, left gage, right gage, and right field), so the total number of images is 813,148. The detection problem consists in determining whether an image contains a fastener attached to one of the rails. The dataset contains bounding boxes for all the images that are known to contain a defect. The total number of defects is 1,087 (0.13% of all the fasteners). The defective fastener class contains two subclasses: broken fastener and missing fastener.

We have used the scores generated by the multi-task learning (MTL) detector described in [3]. This detector uses deep learning with multiple tasks that are trained in parallel. The reason for using multiple tasks is to prevent overfitting. By sharing a common low-level representation between the fastener inspection task and a separate material classification task, there is a data amplification effect that results in better generalization for both classifiers. We also compare the performance with the baseline single-task learning (STL) method in [4]. The raw data was provided by Amtrak, and the authors of [3, 4] provided the output of their detectors as well as the codes to evaluate the performance. This detector produces a scalar-valued score for each image by spatially pooling all the detections in the image. Scores are high when the image contains a good fastener, and low when the fastener is either missing or broken.

Figure 2 shows several detection examples of the MTL detector.

To facilitate the evaluation of fastener detection performance under difficult scenarios, whenever the fastener is not directly attached to the rail or tie, or when for some reason a fastener is not visible at all, those ties are marked as uninspectable with a special label. Depending on the value of such label, the dataset is divided into 3 subsets:

- *Clear ties:* 200,763 ties (1,037 ties with at least one defect).
- *Clear ties plus switches:* 201,856 ties (1,045 ties with at least one defect). See Figure 3 for an example of a switch section.
- *All ties:* 203,287 ties (1,052 ties with at least one defect). This includes switches, and ties for which some fasteners are not visible because they are covered by ballast or a lubricator. See Figures 4 and 5 for examples of high ballast and lubricator sections.

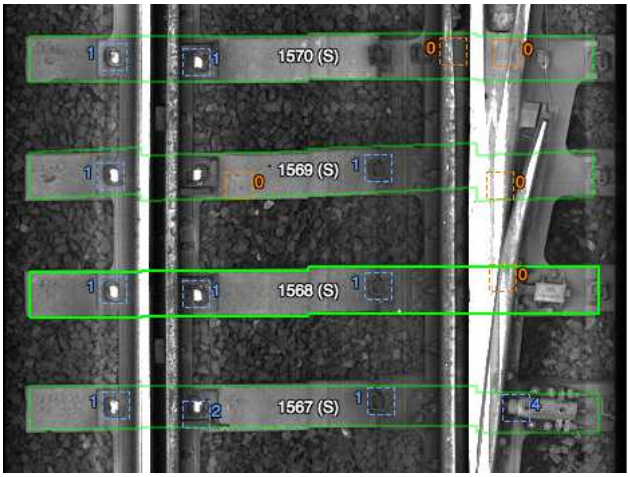


Figure 3. Example of section marked as switch.

For training, we use all the available data after setting aside the sequence being tested. Table 1 and Figure 6 show the detection results on the normalized scores. The overall improvement is significant. The detection rate on the whole dataset at $PFA = 0.1\%$ increases from 95.40% to 99.26%. This is a $6\times$ reduction in the missed rate. Moreover, the performance on the clear tie subset does not degrade at all. The running time of our EVT adaptation algorithm implemented in MATLAB[1] for adapting all 813,148 scores is only of 17 seconds on a Mid-2012 MacBook Pro with a 2.5 GHz Intel Core i5 processor, so this dramatic improvement comes at negligible computational cost (running the detector process takes several hours).

[1]The code and data used in this section is available at https://github.com/xavigibert/EvtTrack

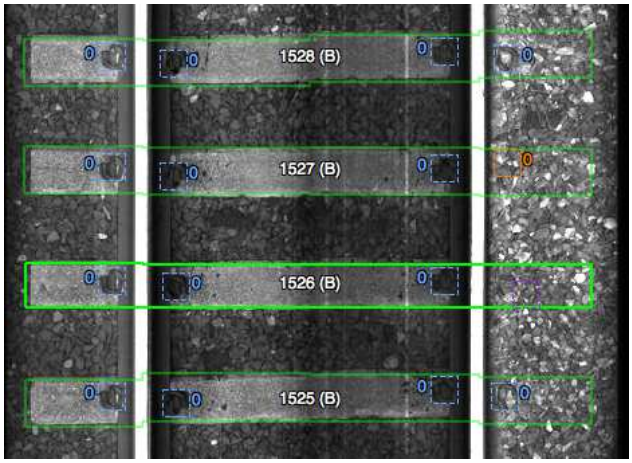


Figure 4. Example of section marked as ballast.

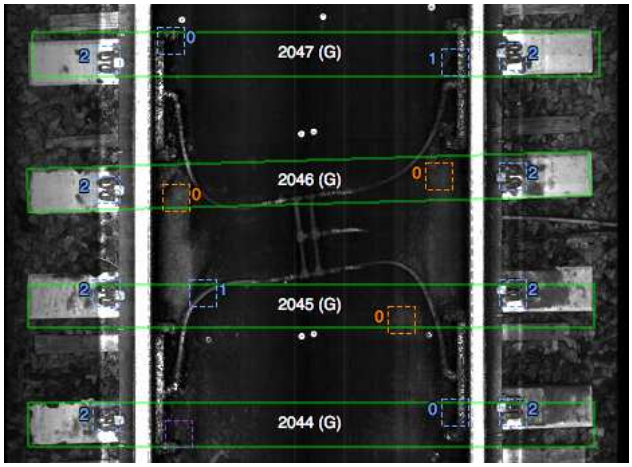


Figure 5. Example of section marked as lubricator.

5. Conclusions

In this paper, we presented a new algorithm that normalizes scores from a sequential anomaly detector with the objective of harmonizing its false alarm rate. Extreme value theory provides a solid foundation from which adaptive thresholding algorithms can be derived. When working with sequences of images, we need to take advantage of the statistical dependencies of nuisance parameters of nearby images. If we discard such dependencies and treat each image in the sequence independently, the performance suffers.

The CFAR detection approach proposed in this paper has applicability beyond railway track inspection from a moving vehicle. It could be used, for example, in surveillance video to remove bursts of false alarms caused by sun glare, insects, rain or fog. Its computational cost is negligible compared to that of the underlying detector, so this approach can be easily retrofitted to existing detectors already

Condition	PFA	MTL + EVT	MTL[3]	STL[4]
Fastener (only clear ties)	0.1% 0.02%	**99.91%** **97.20%**	**99.91%** 96.74%	98.41% 93.19%
Fastener (clear + switch)	0.1% 0.02%	**99.54%** **93.80%**	98.43% 89.35%	94.54% 88.70%
Fastener (all ties)	0.1% 0.02%	**99.26%** **93.47%**	95.40% 87.76%	87.38% –

Table 1. Fastener detection results before and after score normalization.

in operation.

Acknowledgements

The authors thank Amtrak, ENSCO, Inc. and the Federal Railroad Administration for providing the data used in this paper.

References

[1] J. Broadwater and R. Chellappa. Adaptive threshold estimation via extreme value theory. *IEEE Transactions on Signal Processing*, 58(2):490–500, 2010. 3, 5

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[3] X. Gibert, V. M. Patel, and R. Chellappa. Deep multi-task learning for railway track inspection. *arXiv:1509.05267*, 2015. 2, 5, 7

[4] X. Gibert, V. M. Patel, and R. Chellappa. Robust fastener detection for autonomous visual railway track inspection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015. 5, 7

[5] E. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, 1958. 3

[6] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. 2

[7] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, New Jersey, second edition, 2009. 2

[8] D. Y. Kim, J. J. Kim, P. Meer, D. Mintz, and A. Rosenfeld. Robust computer vision: A least median of squares based approach. In *in Proc. of Image Understanding Workshop*, pages 1117–1134, 1989. 2

[9] H. W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967. 5

[10] R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley series in probability and statistics. John Wiley & Sons, Chichester, England, 2006. 2

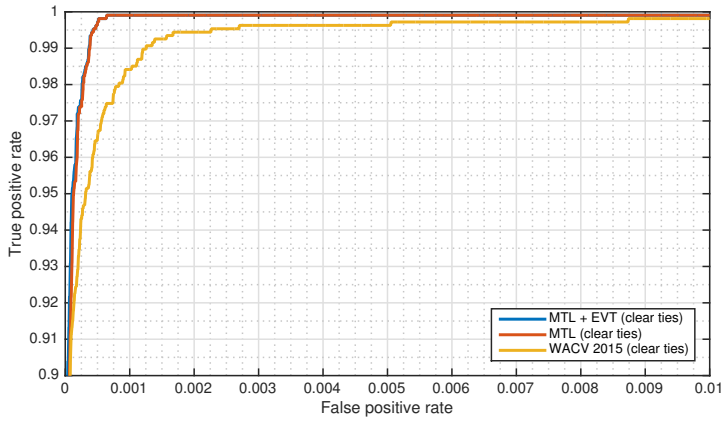
[11] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, jan 1975. 3

[12] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Robust fusion: Extreme value theory for recognition score normalization. In *European Conference on Computer Vision (ECCV)*, pages 481–495. Springer, 2010. 3

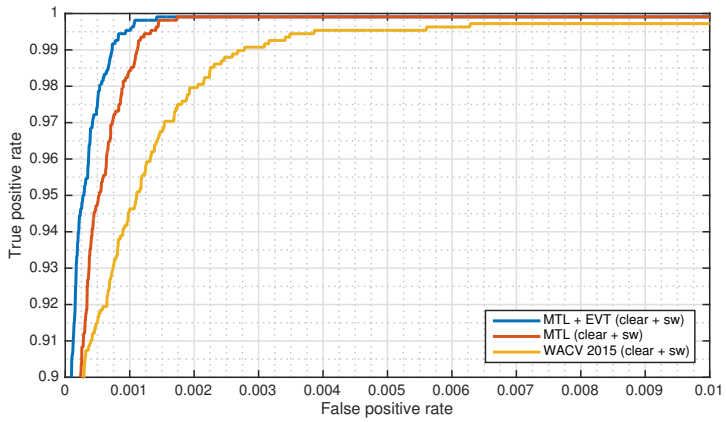
[13] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1689–1695, 2011. 3

[14] A. Stuart, J. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*. Hodder Arnold, London, U.K., 1999. 4

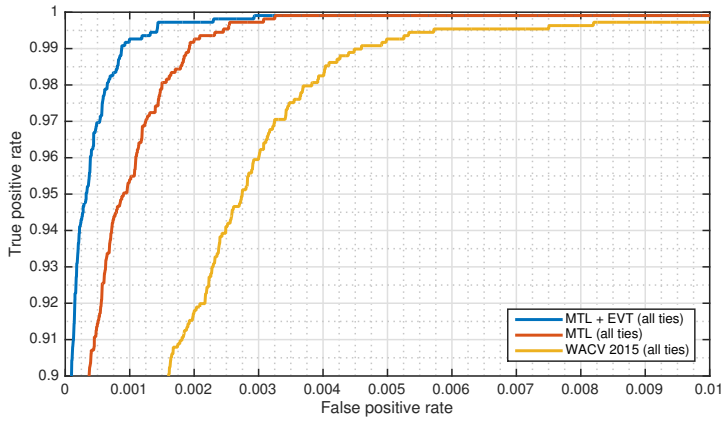
[15] S. Yun. The extremal index of a higher-order stationary markov chain. *The Annals of Applied Probability*, 8(2):408–437, may 1998. 3



(a)



(b)



(c)

Figure 6. ROC curves comparing defective fastener detection performance on the 85-mile testing set using normalized vs. unnormalized scores (a) on the clear ties subset (b) on the clear with with switches subset (c) on all ties. Detections are per image (each tie has 4 images).