On Detection of Median Filtering in Digital Images

Electronic Imaging 2010
Media Forensics and Security II

Matthias Kirchner[†], **Jessica Fridrich**[‡]

[†]Technische Universität Dresden [‡]SUNY Binghamton

San Jose, CA, 2010/01/20

Outline of this Talk

1. Motivation: Detection of Median Filtering?
2. Detection in never-compressed images
3. Detection in JPEG compressed images
4. Conclusion

Digital Image Forensics

- **image forensics:** assess the authenticity of digital images by exploiting their inherent statistical characteristics

source identification

manipulation detection

Digital Image Forensics

- **image forensics:** assess the authenticity of digital images by exploiting their inherent statistical characteristics

source identification

manipulation detection

'malicious' post-processing

'legitimate' post-processing

Digital Image Forensics

- **image forensics:** assess the authenticity of digital images by exploiting their inherent statistical characteristics

source identification

manipulation detection

'malicious' post-processing

- (mostly) local changes
- splicing
- copy & paste
- . . .

'legitimate' post-processing

- content-preserving global changes
- denoising
- compression
- contrast enhancement

definitions depend on established habits and conventions

Digital Image Forensics

- **image forensics:** assess the authenticity of digital images by exploiting their inherent statistical characteristics

source identification

manipulation detection

'malicious' post-processing

'legitimate' post-processing

Digital Image Forensics

- **image forensics:** assess the authenticity of digital images by exploiting their inherent statistical characteristics

source identification

manipulation detection

'malicious' post-processing

'legitimate' post-processing

pictures: Andrea Sommer, Doc Baumann

Processing History of Digital Images

- ’malicious’ post-processing is generally considered to be more critical

but: general processing history of digital images is of great interest

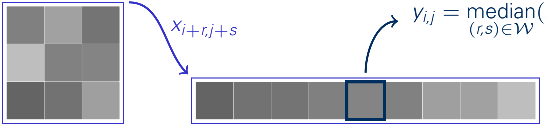
- state of the image prior to the actual (‘malicious’) manipulation may influence
  - the choice of suitable forensic tools
  - the interpretation of results obtained with these tools

  (this applies also to steganalysis) [Böhme, 2009]

- ‘legitimate’ post-processing can interfere with or even wipe out subtle traces of previous manipulations
  - decreased reliability of forensic methods

Detection of Median Filtering

- median filter is a well-known **non-linear** denoising and smoothing operator



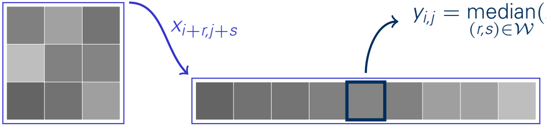
$$y_{i,j} = \underset{(r,s)\in\mathcal{W}}{\operatorname{median}}(x_{i+r,j+s})$$

Why is the detection of median filtering of interest?

- forensic methods often rely on some kind of linearity assumption
  - vulnerable to median filtering [Kirchner & Böhme, 2008]
- smooth(ed) images may require a specific treatment in various applications

Detection of Median Filtering

- median filter is a well-known **non-linear** denoising and smoothing operator



$$y_{i,j} = \operatorname*{median}_{(r,s)\in\mathcal{W}}(x_{i+r,j+s})$$

Why is the detection of median filtering of interest?

- forensic methods often rely on some kind of linearity assumption
  - vulnerable to median filtering [Kirchner & Böhme, 2008]
- smooth(ed) images may require a specific treatment in various applications

Median filtering is hard to model analytically

- highly non-linear and signal-adaptive
- most image processing literature assumes i.i.d. samples

Streaking

- output pixel is drawn directly from the set of input samples
- non-zero probability that output pixels in a certain neighborhood originate from the same input pixel $\longrightarrow$ **streaking** [Bovik, 1987]
- median filtering increases $P_0 = \Pr(y_{i,j} = y_{k,l})$

Streaking

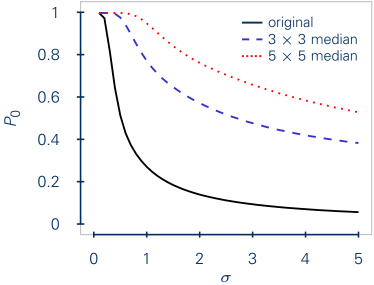
- output pixel is drawn directly from the set of input samples
- non-zero probability that output pixels in a certain neighborhood originate from the same input pixel $\longrightarrow$ **streaking** [Bovik, 1987]
- median filtering increases $P_0 = \Pr(y_{i,j} = y_{k,l})$

indication of median filtering

- for continuous-valued i.i.d. input samples, P_0 is $\boxed{\text{distribution-independent}}$.

Streaking

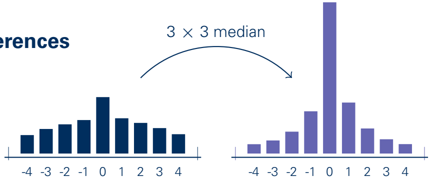
- output pixel is drawn directly from the set of input samples
- non-zero probability that output pixels in a certain neighborhood originate from the same input pixel $\longrightarrow$ **streaking** [Bovik, 1987]
- median filtering increases $P_0 = \Pr(y_{i,j} = y_{k,l})$
- for continuous-valued i.i.d. input samples, P_0 is distribution-independent, but not for discrete signals



streaking probabilities for direct vertical / horizontal neighbors and quantized i.i.d. Gaussian $\mathcal{N}(0, \sigma)$ input samples

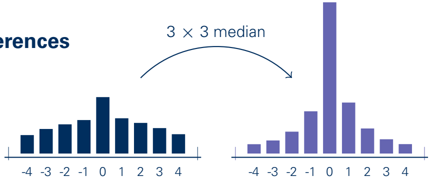
Measuring Streaking Artifacts in Real Images

- histogram of the **first-order differences** $d_{i,j} = y_{i,j} - y_{i+k,j+l}$ with lag (k, l)
- increased peak h_0 due to median filtering

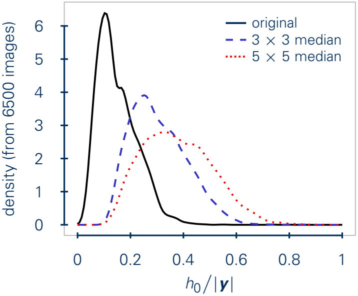


Measuring Streaking Artifacts in Real Images

- histogram of the **first-order differences** $d_{i,j} = y_{i,j} - y_{i+k,j+l}$ with lag (k, l)
- increased peak h_0 due to median filtering

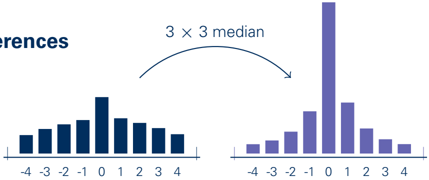


- histogram bin h_0 depends on the image content (smoothness, saturation, ...)

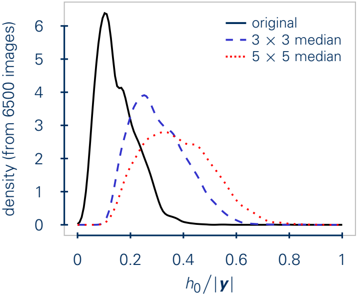


Measuring Streaking Artifacts in Real Images

- histogram of the **first-order differences** $d_{i,j} = y_{i,j} - y_{i+k,j+l}$ with lag (k, l)
- increased peak h_0 due to median filtering



- histogram bin h_0 depends on the image content (smoothness, saturation, ...)

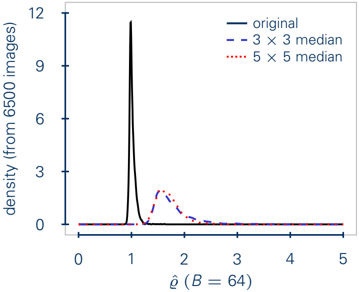


- median filtering increases h_0 relative to h_1
- **normalized measure:** $\varrho = h_0/h_1$
- $\varrho \gg 1$ for median filtered images

Robust Measure

- saturation effects are likely to cause false positives
- assumption: saturation is mostly a localized phenomenon
- measure streaking artifacts in the set $\mathcal{B}$ of all non-overlapping $B \times B$ blocks

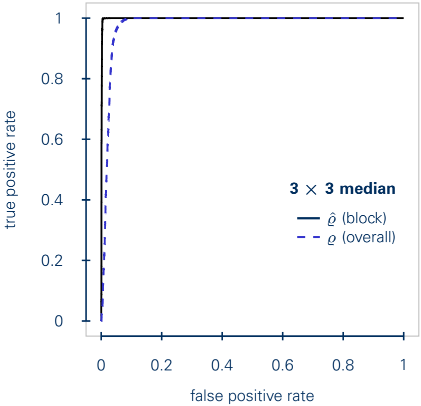
$$\hat{\varrho} = \underset{b \in \mathcal{B}}{\mathrm{median}}(w_b \varrho_b) \qquad \text{with weights} \quad w_b = 1 - \left(\frac{h_0}{B^2 - B} \right)$$



- generally good discrimination between original and filtered images

Experimental Results
overall vs. block-based measure

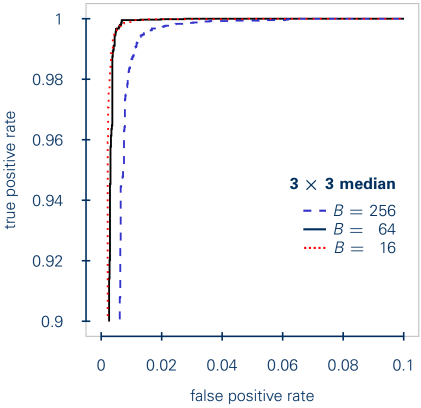
- ▷ database of 6500 images from 22 different cameras
- ▷ never-compressed images, converted to grayscale
- ▷ $(k, l) = (1, 0)$



- ▶ block-based approach $(B = 64)$ is more robust to outliers
- ▶ perfect detection for FPR $< 1.8\,\%$

Experimental Results
influence of block size

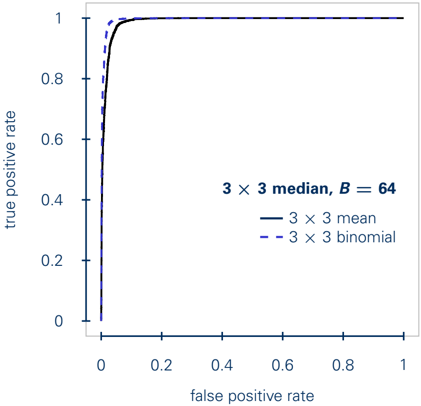
- ▷ database of 6500 images from 22 different cameras
- ▷ never-compressed images, converted to grayscale
- ▷ $(k, l) = (1, 0)$



- ROC curves for block-based approach
- $\hat{\varrho}$ superior for smaller blocks
- too small blocks do not yield additional gain (overall amount of saturation remains the same)
- $B = 64$ suitable choice (for this set of images)

Experimental Results
alternative smoothers

- database of 6500 images from 22 different cameras
- never-compressed images, converted to grayscale
- $(k, l) = (1, 0)$

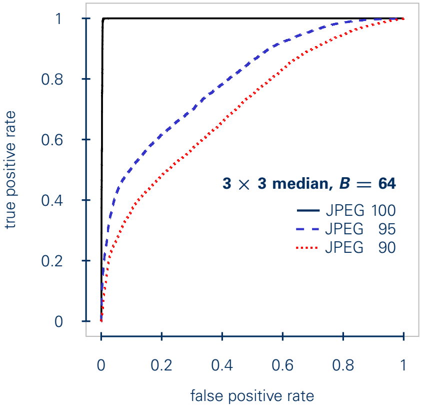


- ROC curves obtained by taking linearly smoothed images as 'originals'
- detector can well distinguish between median filtered and otherwise smoothed images

Experimental Results

JPEG post-compression

- database of 6500 images from 22 different cameras
- never-compressed images, converted to grayscale
- $(k, l) = (1, 0)$

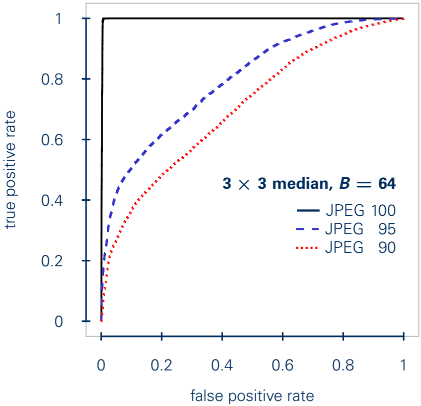


- detector is not robust against JPEG compression

Experimental Results

JPEG post-compression

- ▷ database of 6500 images from 22 different cameras
- ▷ never-compressed images, converted to grayscale
- ▷ $(k, l) = (1, 0)$

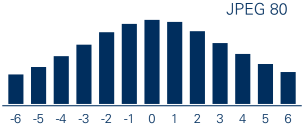


- detector is not robust against JPEG compression
- JPEG smooths the first order differences histogram
- JPEG introduces false alarms

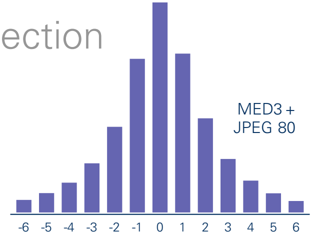
SPAM Features for Median Detection

- smoothing generally affects first-order differences
  - peaky distribution
  - further ‘enhanced’ by subsequent JPEG compression
- strongest effects for small differences $|d_{i,j}| \leq T$

but: generally strong dependence on the image content



SPAM Features for Median Detection



- smoothing generally affects first-order differences
  - peaky distribution
  - further ‘enhanced’ by subsequent JPEG compression
- strongest effects for small differences $|d_{i,j}| \leq T$

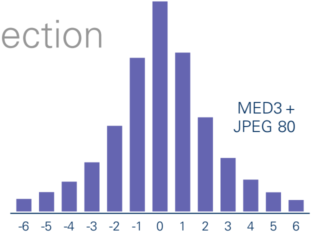
but: generally strong dependence on the image content

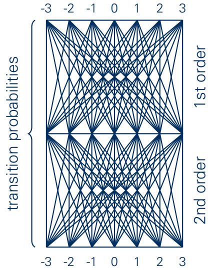
SPAM Features for Median Detection

- smoothing generally affects first-order differences
  - peaky distribution
  - further ‘enhanced’ by subsequent JPEG compression
- strongest effects for small differences $|d_{i,j}| \leq T$

but: generally strong dependence on the image content

- more sophisticated model: **SPAM features** [Pevný et al., MM-Sec 2009]
- subtractive pixel adjacency matrix models first-order differences as n-th order Markov chain
- transition probabilities (= conditional joint distribution) taken as features in a high-dimensional classification problem





SPAM Details

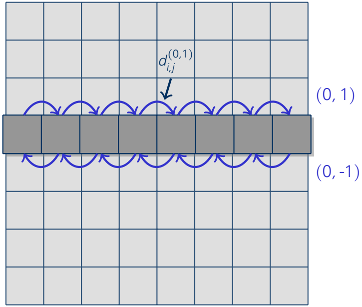
- transition probabilities for first-order differences $d_{i,j}^{(k,l)}$ with lag $(k,l) \in \{\text{-}1, 0, 1\}^2$

$$M_{\delta_n,\ldots,\delta_0}^{(k,l)} = P\left(d_{i+kn,j+ln}^{(k,l)} = \delta_n \,\middle|\, d_{i+k(n-1),j+l(n-1)}^{(k,l)} = \delta_{n-1}, \ldots, d_{i,j}^{(k,l)} = \delta_0\right)$$

SPAM Details

- transition probabilities for first-order differences $d_{i,j}^{(k,l)}$ with lag $(k,l) \in \{\text{-}1, 0, 1\}^2$

$$M^{(k,l)}_{\delta_n,\ldots,\delta_0} = P\left(d^{(k,l)}_{i+kn,j+ln} = \delta_n \,\middle|\, d^{(k,l)}_{i+k(n-1),j+l(n-1)} = \delta_{n-1}, \ldots, d^{(k,l)}_{i,j} = \delta_0\right)$$



- horizontal / vertical transition matrices

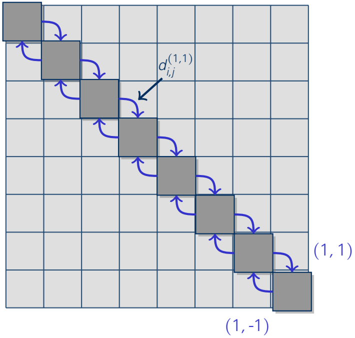
$$\mathbf{M}^{(0,1)} \quad \mathbf{M}^{(0,\text{-}1)}$$

$$\mathbf{M}^{(1,0)} \quad \mathbf{M}^{(\text{-}1,0)}$$

SPAM Details

- transition probabilities for first-order differences $d_{i,j}^{(k,l)}$ with lag $(k,l) \in \{\text{-}1,0,1\}^2$

$$M_{\delta_n,\ldots,\delta_0}^{(k,l)} = P\left(d_{i+kn,j+ln}^{(k,l)} = \delta_n \,\middle|\, d_{i+k(n-1),j+l(n-1)}^{(k,l)} = \delta_{n-1},\ \ldots\ , d_{i,j}^{(k,l)} = \ \delta_0\right)$$



- horizontal / vertical transition matrices

$$\mathbf{M}^{(0,1)} \quad \mathbf{M}^{(0,\text{-}1)}$$

$$\mathbf{M}^{(1,0)} \quad \mathbf{M}^{(\text{-}1,0)}$$

- diagonal transition matrices

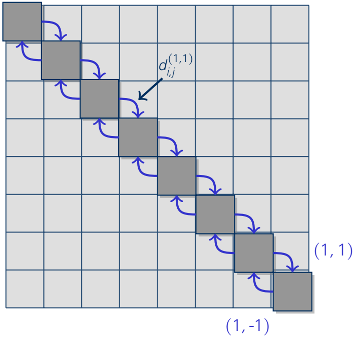
$$\mathbf{M}^{(1,1)} \quad \mathbf{M}^{(1,\text{-}1)}$$

$$\mathbf{M}^{(\text{-}1,\text{-}1)} \quad \mathbf{M}^{(\text{-}1,1)}$$

SPAM Details

- transition probabilities for first-order differences $d_{i,j}^{(k,l)}$ with lag $(k,l) \in \{\text{-}1, 0, 1\}^2$

$$M_{\delta_n,\ldots,\delta_0}^{(k,l)} = P\left(d_{i+kn,j+ln}^{(k,l)} = \delta_n \,\middle|\, d_{i+k(n-1),j+l(n-1)}^{(k,l)} = \delta_{n-1}, \ldots, d_{i,j}^{(k,l)} = \delta_0\right)$$



- horizontal / vertical features

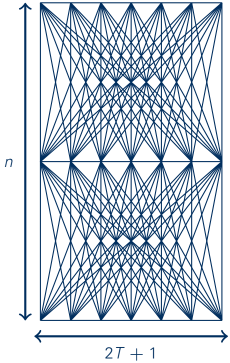
$$\mathbf{F}^{(h/v)} = 1/4\Big(\mathbf{M}^{(0,1)} + \mathbf{M}^{(0,\text{-}1)} + \mathbf{M}^{(1,0)} + \mathbf{M}^{(\text{-}1,0)}\Big)$$

- diagonal features

$$\mathbf{F}^{(d)} = 1/4\Big(\mathbf{M}^{(1,1)} + \mathbf{M}^{(1,\text{-}1)} + \mathbf{M}^{(\text{-}1,\text{-}1)} + \mathbf{M}^{(\text{-}1,1)}\Big)$$

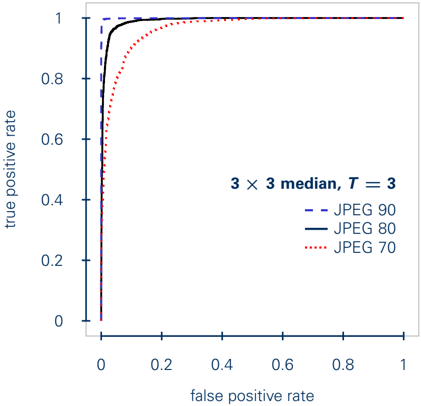
SPAM Classifier

- **number of features:** $2\,(2T+1)^{n+1}$
- in our tests: $n = 2$ and $T \in \{1, 2, 3\}$
  - up to 686 features
- soft-margin SVM with Gaussian kernel
  - one classifier per filter size and JPEG post-compression quality
  - parameter search and training with ≈ 3250 images per class (five-fold cross-validation)
  - validation with another ≈ 3250 images per class



Experimental Results
SPAM features

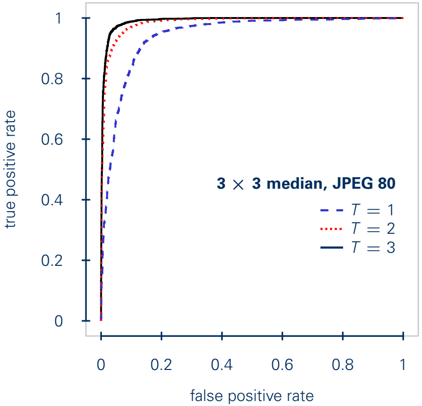
- database of 6500 images from 22 different cameras
- never-compressed images, converted to grayscale
- 512×512 center region



- high detectability even for rather strong JPEG compression

Experimental Results
SPAM features

- ▷ database of 6500 images from 22 different cameras
- ▷ never-compressed images, converted to grayscale
- ▷ 512×512 center region

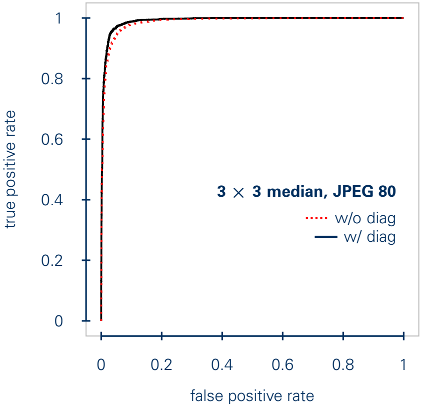


- high detectability even for rather strong JPEG compression
- higher SPAM dimensionality increases performance

Experimental Results

SPAM features

- database of 6500 images from 22 different cameras
- never-compressed images, converted to grayscale
- 512 × 512 center region

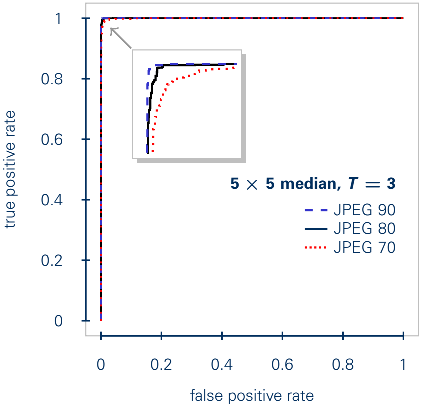


- high detectability even for rather strong JPEG compression
- higher SPAM dimensionality increases performance
- diagonal features do not provide additional information beyond horizontal / vertical features

Experimental Results

SPAM features

- database of 6500 images from 22 different cameras
- never-compressed images, converted to grayscale
- 512 × 512 center region



- high detectability even for rather strong JPEG compression
- higher SPAM dimensionality increases performance
- diagonal features do not provide additional information beyond horizontal / vertical features
- considerably improved performance for larger filter sizes

Further Experiments

- lower-order Markov models yield slightly worse results
- larger images result in better performance
- pre-median JPEG compression does not seem to influence detection results

Further Experiments

- lower-order Markov models yield slightly worse results
- larger images result in better performance
- pre-median JPEG compression does not seem to influence detection results
- SPAM features **cannot** distinguish between median filter and other smoothers
  - similar effects w. r. t. the distribution of small first-order differences
  - **SPAM as a general-purpose smoothing detector?**

Concluding Remarks

- general processing history is of great interest in various situations
  - make **informed decisions** in image forensics, steganalysis and watermarking
- in this study: detection of median filtering

first-order differences

streaking artifacts
zeroth-order Markov chain
$\varrho = P_{\delta_0} / P_{\delta_1}$

smoothness
n-th order Markov chain
$F = P_{\delta_n,\ldots,\delta_0}$

- JPEG post-compression obfuscates the actual type of smoothing
  - SPAM as general-purpose detector
  - explore alternative / additional features that are more specific to median filtering

Thanks for your attention

Questions?

Matthias Kirchner[†], **Jessica Fridrich**[‡]

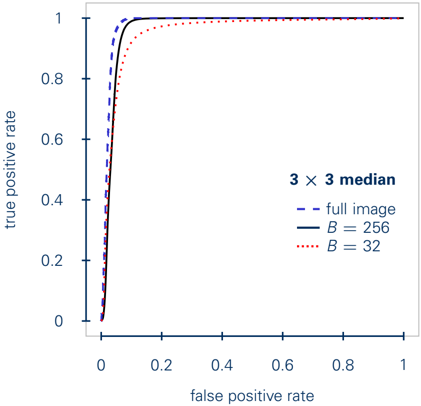
[†]Technische Universität Dresden [‡]SUNY Binghamton

Matthias Kirchner gratefully receives a doctorate scholarship from Deutsche Telekom Stiftung, Bonn, Germany.

Experimental Results

per-block decision

- ▷ database of 6500 images from 22 different cameras
- ▷ never-compressed images, converted to grayscale
- ▷ $(k, l) = (1, 0)$



- ROC curves over all non-overlapping blocks of all images
- ϱ_b itself is more sensitive to local variations throughout the image
- larger blocks are beneficial for a per-block decision (local detection of median filtering)