Watermark detection for video bookmarking using mobile phone camera

Peter Meerwald* and Andreas Uhl

Dept. of Computer Sciences, University of Salzburg,
Jakob-Haringer-Str. 2, A-5020 Salzburg, Austria
{pmeerw, uhl}@cosy.sbg.ac.at, http://www.wavelab.at

Abstract. In this paper we investigate a watermarking application for bookmarking of video content using a mobile phone's camera. A content identifier and time-stamp information are embedded in individual video frames and decoded from a single frame captured from a display device, allowing to remember ('bookmark') scenes in the video. We propose a simple watermarking scheme and blind image registration to combat the inherent geometric distortion due to digital/analog conversion. The work-in-progress shows promising results over previous approaches.

Key words: Watermarking, image registration, geometric distortion

1 Introduction

Watermarking has been proposed as a technology to embed an imperceptible, yet detectable signal in digital multimedia content such as images or video [1]. Since the watermark information is embedded in the video data itself and not a particular file format, the embedded information is retained even if the content undergoes transformation such as re-encoding or presentation on a monitor and capturing with a camera. Pramila et al. [2] survey the challenges in bridging the analog/digital gap using camera-based watermark extraction. Transmission of watermark information over the print/scan channel has been studied for applications including document authentication and copyright protection [3–5].

Nakamura et al. [6, 7] present two watermark detection schemes for camera-equipped cellular phones: in [6], the authors aim to decode information such as an imperceptible content identifier analogous to a visible bar code in printed material while in [7], a content id is decoded from a sequence of video frames. Both methods rely on extraction of the target content region in the captured image and image sequence before applying projective correction to combat geometric distortion which inevitably results from freehand shooting. The side trace algorithm (STA) [8] employed requires a smooth background or an artificial border marker in order to identify the target region.

Stach et al. investigate the use of web cameras for watermark detection [9], most of the challenges identified (demosaicking, lens distortion, focus issues,

* Supported by Austrian Science Fund project FWF-P19159-N13.

white balance and gain compensation, compression) also apply to mobile phone cameras.

Watermarking robust to complex geometric distortion is a requirement for the digital cinema application [10–12] where the watermark may reveal the particular cinema in which in camcorder copy ('screener') was made: The early work [11] requires the original movie for synchronization, the later proposal [12] spread the watermark information over large temporal regions and thus cannot pinpoint individual frames; both approaches do not satisfy real-time detection requirement. Lee et al. [10] propose to use the the same watermark for information embedding and for geometric distortion estimation via local auto-correlation function; however, they require high-resolution video and a controlled capture environment.

In this work we investigate a watermarking application where a content identifier and time-stamp information are embedded in individual video frames and decoded from a single frame captured from a display device. This allows to remember ('bookmark') scenes in the video by means of decoding the embedded time-stamp and content id information. We discuss the application requirements and trade-offs in Section 2 and propose a simple watermarking scheme in Section 3. The blind image registration procedure to combat the inherent geometric distortion is presented in Section 4. Performance evaluation and experimental results are provided in Section 5, followed by concluding remarks in Section 6.

2 Application Scenario

In the envisioned application, a content identifier and time-stamp information are embedded in individual video frames. By taking pictures of video scenes displayed on a monitor with a mobile phone's camera, bookmarks can be conveniently stored as the embedded time-stamp information in the captured pictures links back to the particular scenes of the video content.

Hirakawa et al. [13] study the use of watermarking in conjunction with mobile devices and put forward marketing as a promising field of application instead of the well-researched copyright protection scenario. Nakamura et al. [6, 7] propose a so-called Related Service Introduction System (RSIS) which enables auxiliary program information, online shopping, etc., based on the content identifier transmitted with the content itself.

For the video bookmark application, watermarking can be used to imperceptibly and unobtrusively store the time-stamp information in the video and thus cross the media boundary. Instead of watermarking, perceptual hashing could be used as a passive alternative, i.e. without having to modify the video content to embed the watermark information. Both technologies have their merits: (i) Watermark decoding can be performed stand-alone by the mobile device. This allows to give feedback to the user immediately, e.g. when capture conditions do not allow reliable detection. The embedded information can be extracted by the mobile device or the captured pictures can be uploaded to a server that does the processing. Clearly, there is a bandwidth versus processing trade-off. Perceptual

hashing requires a query to a database server to turn the perceptual hash into meaningful information. (ii) In case of active techniques, content preparation can be time-consuming due to the application of the watermark; furthermore, there is always at least a slight quality degradation due to watermark embedding. Perceptual hashing only requires to compute the hashes, without altering the content. (iii) Perceptual hashing cannot distinguish the same content coming from two different sources. On the other hand, by embedding different watermarks, one can link different services to the same content distributed via DVD or TV broadcast for example [7].

For the given video bookmarking application, 56 bits of watermark capacity is necessary to transmit 32 bits content id and 24 bits time-stamp information which permits temporal addressing of more than three days of video at 60 frames/second. A high video quality should be maintained; in this work we aim for 43 dB PSNR; this is considerable higher than the PSNR achieved by previous single frame watermark detection schemes (37 dB PSNR [6]), but lower than the quality obtained by methods utilizing multiple frames for detection (48 dB PSNR [10]; 49 dB PSNR [7]). Efficient processing is necessary for real-time application but we do not attempt to implement the algorithms on a mobile platform at this time.

3 Watermark Embedding and Detection

The watermark payload vector $\mathbf{b}$ is embedded in the luminance component of the video frames of size $N \times M$ pixels. $\mathbf{b}$ consists of $B = 56$ bits, a concatenation of the 32 bit content id and 24 bits time-stamp information identifying the temporal frame position. Payload bits are denoted $b[i] \in \{-1, 1\}$ with $1 \leq i \leq B$.

We perform a two-level discrete wavelet transform (DWT) with biorthogonal 7/9 filters and concatenate the resulting detail subband coefficients of the HL_2, LH_2, and HH_2 subband into the host signal vector denoted by $\mathbf{x}$. Next, we permute the coefficients of $\mathbf{x}$ and partition the resulting vector into B non-overlapping blocks of equal size, the coefficients of each block are denoted $x_i[k]$ with $1 \leq i \leq B$, $1 \leq k \leq S$ and $S = {}^{3 \cdot N \cdot M}/_{4 \cdot B}$. The permutation guarantees that the watermark information contributing to one payload bit is spread out equally over the host signal.

Embedding of the payload in the detail subbands using additive, spread-spectrum watermarking can be written

$$x'_i[k] = x_i[k] + \alpha \cdot b[i] \cdot w_i[k] \tag{1}$$

where α is the embedding strength. The spreading sequence $\mathbf{w}$ is generated pseudo-randomly with symbols $w[k] \in \{-1, 1\}$ of equal probability. No perceptual shaping is performed in this work, we simply set $\alpha = 3$.

3.1 Watermark Detection and Decoding

Watermark detection is performed *blind*, i.e. without reference to the original host signal. For efficient blind watermark detection, accurate modeling of the

host signal is required. We assume a Cauchy distribution of the DWT details subband coefficients and chose the Rao-Cauchy (RC) detector [14] whose detection statistic for the received signal $\mathbf{y}$ of length L is given by

$$\rho = \frac{8\hat{\gamma}^2}{L} \left[\sum_{t=1}^{L} \frac{y[t] \cdot w[t]}{\hat{\gamma}^2 + y[t]^2} \right]^2 \tag{2}$$

where γ is an estimate of the Cauchy distribution's scale parameter. To decide between the null- ($\mathcal{H}_0$, no or other watermark present) and alternative hypothesis ($\mathcal{H}_1$, watermark $\mathbf{w}$ present), the detection statistic ρ is compared against a threshold, $\rho \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} T$. The detection statistic ρ follows a Chi-Square distribution with one degree of freedom (χ_1^2) under $\mathcal{H}_0$ and, under $\mathcal{H}_1$, a non-central Chi-Square distribution with non-centrality parameter λ ($\chi_{1,\lambda}^2$) [15]. The detection threshold T can be established in the Neyman-Pearson sense for a desired probability of false-alarm using the relation $P_f = \mathbb{P}\{\rho > T|\mathcal{H}_0\} = Q_{\chi_1^2}(T)$ and the identity $Q_{\chi_1^2}(x) = 2Q(\sqrt{x})$ where the function $Q(\cdot)$ expresses the right-tail probability of the Gaussian distribution [16]. It follows that $T = \left[Q^{-1}(P_f/2)\right]^2$. For estimation of the Cauchy scale parameter γ, several fast, approximate methods are available, e.g. [17].

The RC detector has the advantage that the detection threshold does not depend on the received signal and can be precomputed to meet the desired false-alarm probability. Further, knowledge of the watermark strength is not necessary for detection [15]. These properties are important in the investigated application scenario as extensive experiments for threshold determination would be time-consuming to perform and the watermark power is in most instances significantly impaired due to the signal's analog/digital transition.

In order to retrieve the embedded payload, the target video content must be located and extracted from a captured image. Then the geometric distortion has to be estimated and reversed in a projective correction step. We discuss this blind registration process in the next section. For now, assume that we have a candidate received frame the same size as the original frame which is subjected to the DWT and used to construct the received signal vector $\mathbf{y}$ the same ways $\mathbf{x}$ was derived earlier.

First, we try to establish the presence or absence of any watermark information on the full host signal vector $\mathbf{y}$. Only if successful, the payload is decoded from the individual blocks $\mathbf{y}_i$. On absence, different parameters can be tried in the frame registration process described in Section 4 or in the shift compensation process, see Fig. 1.

For watermark detection, we remember that the watermark sequence $\mathbf{w}$ is known, but not the payload $\mathbf{b}$. Due to the square operation, the RC detection statistic for each block (Eq. 2) does not depend on the embedded payload bit. Assuming that the detector response of each block $\mathbf{y}_i$ is independent, we can exploit the reproductivity property of the Chi-Square distribution, namely that the sum of Chi-Square random variables is again Chi-Square distributed. We have B random variables $\rho_1, \ldots, \rho_B$, which all follow Chi-Square distributions

with one degree of freedom, thus the sum follows a Chi-Square distribution with B degrees of freedom, $\sum_{i=1}^{B} \rho_i \sim \chi^2_B$, and the combined threshold can be determined using the inverse of the Chi-Square cumulative distribution function with B degrees of freedom.

For payload decoding, we simply compute the sign of the modified detection statistic (without the square operation) for each received signal block

$$b[i] = \operatorname{sgn}\left(\sum_{k=1}^{S} \frac{y_i[k] \cdot w_i[k]}{\hat{\gamma_i}^2 + y_i[k]^2}\right) \tag{3}$$

which is equivalent to a bit-by-bit hard-decision decoder. The proposed decoder uses binary, antipodal constellation of codewords and space-division multiplexing.

In the proposed approach, no pilot or template watermark is used to estimate distortion, hence the entire watermark energy can be spent on encoding the payload. Compared to very recent work [10], a more sophisticated but still computationally efficient host signal model and corresponding detector is employed. Further, the spread-spectrum scheme is implicitly robust against certain valumetric distortion, in particular contrast change, which is an important advantage over quantization-based schemes [1].

4 Blind Image Registration

As a result of camera capture, the rectangular video frame is transformed into a quadrangle due to perspective projection [10]. For correlation-based detectors, the received signal and the reference watermark must be properly aligned (synchronized) for successful detection and decoding. The strategy we adopt in this work is to estimate and invert distortions in the detector, without reference to the original signal. Note that in our application scenario, no malicious attacker aims to thwart the decoding process by malicious transformation; on the contrary, the user may be asked to improve capture conditions or aid in identification of the target video frame.

The overall process is depicted in Fig. 1. Given a captured image, the first task is to locate the candidate target frame. We assume that the entire video frame has been captured and consumes a significant, large area in the image. Fig. 2 illustrates the processing steps for locating candidate target frames.

In the second task, a candidate target frame is then subjected to a corrective projection in order to invert the distortion. Given the coordinates of the target frame quadrangle, this transform can be easily computed, e.g. using ImageMagick's[1] *distort perspective* feature. The output of the corrective projection is a registered video frame, the same size as the original frame.

Finally, the registered, received frame is fed to the geometric shift compensation unit and the watermark detector. Shift compensation simply applies a small number of integer pixel shifts in each dimension before applying the DWT; this

[1] ImageMagick is available at `http://www.imagemagick.org`.

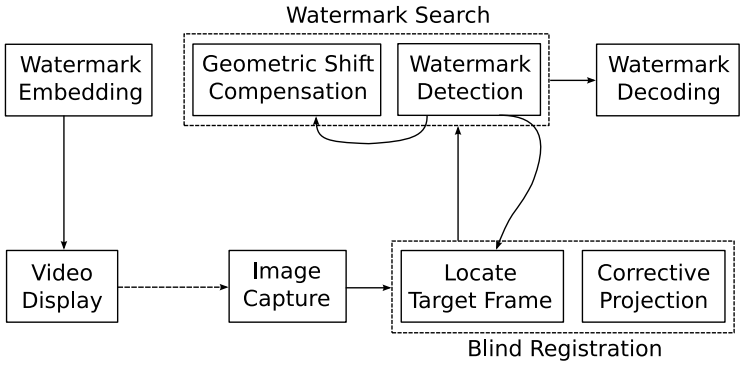


Fig. 1: Overview of video bookmarking system's components.

step is necessary since the (decimated) DWT is not shift-invariant and produces drastically different results on shifted version of the same input.

4.1 Target Frame Localization

The goal is to produce a list of coordinate tuples (p'_1, p'_2, p'_3, p'_4) identifying the corner points of suspect video frames. To this end we use the contrast between the target video frame and the background to identify edges and, subsequently, corners of the target frame. As a pre-processing step, the captured image is Gaussian filtered (7×7 kernel) to suppress image noise. Then, Canny edge detection is performed to produce a gradient magnitudes image which serves as the input for a probabilistic Hough transform. The Hough transform computes a list of the dominant line segments from which approximately horizontally or vertically oriented lines are selected. We assume that geometry of the original video frame in only mildly distorted, i.e. border lines are rotated less than 10 degrees. The endpoints of the selected line segments are used as the center location for a local (33×33 pixels) template matching operation in the gradient magnitude image with the aim to pinpoint the corner coordinates. The templates are small, idealized corner images in four orientations. Note that the line segments already provide orientation information (left or right corner for a horizontal line, top or bottom corner for a vertical line) which is used in the template matching step; the template matching step removes the remaining ambiguity and allows to associate the point of peak normed correlation with one of the target frame coordinate, $p'_1, \ldots, p'_4$. If the maximum normed correlation computed by template matching is below a threshold (< 0.7), the line endpoint is discarded. In Fig. 4, a gradient magnitudes image is shown with line segments superimposed. Small circles denote line endpoints, large circles indicate potential frame corners.

Compared to a previous approach (STA, [8]), the proposed method does not require an explicit border marker, nor a uniform background for target frame localization. Nevertheless, blind localization of the target video frame is probably

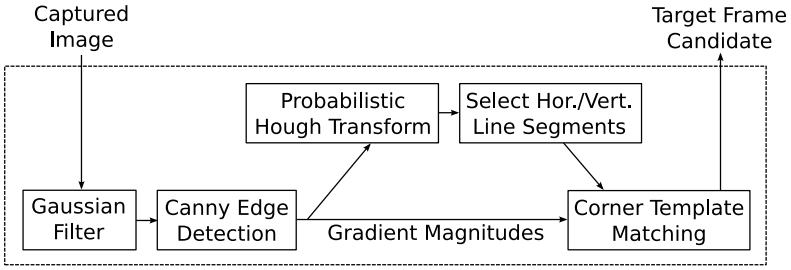


Fig. 2: Target frame location process within the blind image registration component.

the Archilles' heel of the scheme; given the nature of the application, it might be acceptable to ask the user for assistance in a semi-automated way.

5 Experimental Results

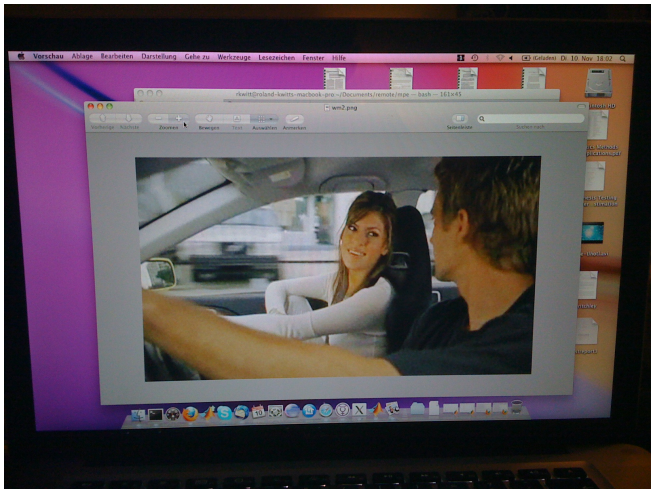
In order to assess the performance of the proposed watermarking system, we select two 480×272 pixel frames of the *2 Fast and 2 Furious* trailer video sequence. The watermark is embedded only in the luminance component to encode 56 bits of payload; the embedding strength $\alpha = 3$ results in an average PSNR of ~ 43 dB. In Fig. 3a we show the watermarked frame #2. The image captured using the camera of an Apple iPhone 3GS (1600×1200 pixels) is presented in Fig 3b.

A critical step is the localization of the target video frame in the captured image. The chosen picture, Fig. 3b, provides an interesting example as several quadrangles can be identified. In Fig. 4 we show the gradient magnitude image produced by the Canny edge detector (cf. Fig. 2). Detected horizontal and vertical line segments have been superimposed (in *red*). Small *(blue)* circles denote line segment endpoints while larger *(green)* circles identify the corner points of candidate quadrangular target frames and result from a positive corner template matching step. It can be seen that the corners of the target video frame are among the potential corners identified.

In Fig. 3c we show the correctly identified 480×272 target video frame extracted using the coordination from the localization step above, after applying corrective projection using ImageMagick. As becoming evident from the light-gray stripe at the top of the extracted frame, the identified top left corner has not precisely located the corresponding quadrangle's coordinate. This imperfection is mitigated in the watermark search step in the spatial (integer pixel offsets $0, \ldots, 3$) and wavelet domain (integer coefficient offsets $-1, \ldots, 1$).

In Table 1 we present the watermark detection results (ρ) and the achieved bit error rate (BER) for the decoded payload on two video frames. The images have been obtained with different mobile phone camera models (and one digital camera for comparison) covering a wide range of resolutions and picture quali-

(a) Watermarked 480×272 pixels video frame (43 dB PSNR).



(b) Image captured with iPhone camera, 1600×1200 pixels.

(c) Extracted 480×272 image after corrective projection.

Fig. 3: Example watermarked video frame.

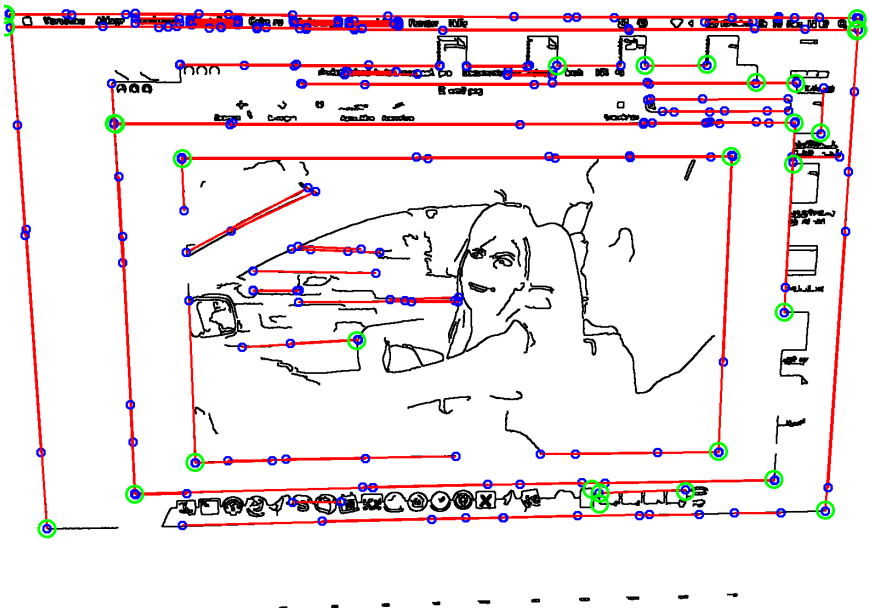


Fig. 4: Gradient magnitude image with line segments and candidate target frame corners superimposed.

ties. No specific instructions were given regarding the presentation of the video frames; all captured images resulted from free-hand shooting.

For a false-alarm probability of $P_f = 10^{-9}$, the detection threshold $T \approx 144.30$ can be determined (assuming $\rho \sim \chi^2_{56}$ under $\mathcal{H}_0$). We observe that the detection statistic is above the detection threshold in all cases. In most cases, the 56 bit payload can be perfectly recovered. Even with camera resolution as low as 640×480 times, watermark recovery is possible. On the other hand, a high-resolution sensor does not guarantee error-free decoding in unfavorable capture conditions (lighting, small target frame size) as is the case for the Nokia 6303 images. All captured images have been JPEG compressed by the camera with default quality settings producing image files between over 500 KByte in case of the iPhone, or as small as 40 KByte in case of older Samsung and Sony models. Results marked with * have been obtained by manually specifying one of the coordinates of the target video frame, i.e. the automatic localization has failed.

All source code is available at `http://www.wavelab.at/sources`. Watermark embedding and detection/decoding has been implemented in Python using the `numpy`/`scipy` libraries[2], the target frame localization code builds on the OpenCV library[3]. The code is not optimized for performance. Watermark de-

[2] See `http://www.numpy.org` and `http://www.scipy.org`.
[3] Available at `http://opencv.willowgarage.com`.

Phone / Camera Model and Resolution	Frame #1		Frame #2	
	ρ	BER	ρ	BER
Apple iPhone 3GS (1600 × 1200)	349.25	0.00%	901.07	0.00%
Motorola Razr v3 (640 × 480)	202.37	3.57%	204.10	5.36%
Nokia 6300i (1600 × 1200)	710.80	0.00%	1211.41	0.00%
Nokia 6303 (2048 × 1536)	*361.94	*1.79%	*213.96	*5.36%
Nokia E51 (640 × 480)	431.71	0.00%	1524.15	0.00%
Nokia N97 (2592 × 1944)	909.52	0.00%	1972.83	0.00%
Qtek 2020i (640 × 480)	320.28	0.00%	524.51	0.00%
Samsung SGH-X550 (640 × 480)	*204.77	*1.79%	220.37	1.56%
Samsung SGH-F480 (2560 × 1920)	452.11	0.00%	2483.59	0.00%
Sony Ericsson W200 (640 × 480)	482.55	0.00%	505.30	0.00%
Sony Ericsson W302 (1600 × 1200)	*162.27	*5.36%	667.95	0.00%
Sony Ericsson K550i (1632 × 1224)	*312.67	*0.00%	395.61	0.00%
Canon IXUS 70 (3072 × 2304)	586.16	0.00%	972.85	0.00%

Table 1: Detection (ρ) and decoding results (BER) for two watermarked video frames (480 × 272) captured with different cameras; payload 56 bits, embedding with 43 dB PSNR.

tection including all shift compensation steps takes approximately half a second on a Intel Core2 2.33 GHz. Blind registration consumes less than half a second to compute the target frame coordinates and approximately 700 ms for each perspective correction attempt.

6 Conclusion

We presented a watermarking application for bookmarking of video scenes by decoding time-stamp and content identifier information from captures frames using mobile phone cameras. We motivated our design decisions by comparing with recent related work. The unique requirement of the application to provide watermark information with high temporal resolution in the video content, and the monitor/camera capture transmission channel have not been studied before.

Universal Pictures Germany GmbH acknowledged that permission is not required to use screen shots taken of the *2 Fast 2 Furious* trailer for academic purpose. The author would like to thank all contributors who provided their mobile phones' pictures.

References

1. Cox, I.J., Miller, M.L., Bloom, J.A., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography. Morgan Kaufmann (2007)
2. Pramila, A., Keskinarkaus, A., Seppänen, T.: Camera based watermark extraction - problems and examples. In: Proceedings of the Finnish Signal Processing Symposium 2007, Oulu, Finland (August 2007)
3. He, D., Sun, Q.: Practical print-scan resilient watermarking scheme. In: Proceedings of the IEEE International Conference on Image Processing, ICIP '05. Volume 1., Genova, Italy, IEEE (September 2005) 257–260
4. Solanki, K., Madhow, U., Manjunath, B.S., Chandrasekaran, S., El-Khalil, I.: 'print and scan' resilient data hiding in images. IEEE Transactions on Information Forensics and Security **1**(4) (August 2006) 464–478
5. Kim, W.G., Lee, S.H., Seo, Y.S.: Image fingerprinting scheme for print-and-capture mode. In: Advances in Multimedia Information Processing, Proceedings of the Pacific Rim Conference on Multimedia, PCM '06. Volume 4261 of Lecture Notes in Computer Science., Springer (2006) 106–113
6. Nakamura, T., Katayama, A., Yamamuro, M., Sonehara, N.: Fast watermark detection scheme for camera-equipped cellular phone. In: Proceedings of the 3rd International Conference on Mobile and ubiquitous Multimedia, College Park, MD, USA, ACM (October 2004) 101–108
7. Nakamura, T., Yamamoto, S., Kitahara, R., Katayama, A., Yasuno, T., Sonehara, N.: A fast, robust watermark detection scheme for videos captured on camera phones. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME '07. (July 2007) 316–319
8. Katayama, A., Nakamura, T., Yamamuro, M., Sonehara, N.: New high-speed frame detection method: Side trace algorithm (STA) for i-appli on cellular phones to detect watermarks. In: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia, College Park, MD, USA, ACM (October 2004) 109–116
9. Stach, J., Brundage, T.J., Kirk, T., Bradley, B.A., Brunk, H.: Use of web cameras for watermark detection. In Delp, E.J., Wong, P.W., eds.: Proceedings of SPIE, Security and Watermarking of Multimedia Contents IV. Volume 4675., San Jose, CA, USA, SPIE (January 2002) 611–620
10. Lee, M.J., Kim, K.S., Suh, Y.H., Lee, H.K.: Improved watermark detection robust to camcorder capture based on quadrangle estimation. In: Proceedings of the IEEE Internationel Conference on Image Processing, ICIP '09, Cairo, Egypt, IEEE (November 2009) 101–104
11. Delannay, D., Delaigle, J.F., Macq, B.M., Barlaud, M.: Compensation of geometrical deformations for watermark extraction in the digital cinema applications. In: Proceedings of SPIE, Security and Watermarking of Multimedia Contents III. Volume 4314., San Jose, CA, USA (January 2001) 149–157
12. Lubin, J., Bloom, J., Hui, C.: Robust, content-dependent, high-fidelity watermark for tracking in digital cinema. In: Proceedings of SPIE, Security and watermarking of multimedia contents V. Volume 5020., San Jose, CA, USA, SPIE (January 2003) 536–545
13. Hirakawa, M., Iijima, J.: A study on digital watermarking usage in the mobile marketing field: Cases in Japan. In: Proceedings of the 2nd International Symposium on Logistics and Industrial Informatics, LINDI '09, Linz, Austria (September 2009) 154–159

14. Kwitt, R., Meerwald, P., Uhl, A.: A lightweight Rao-Cauchy detector for additive watermarking in the DWT-domain. In: Proceedings of the ACM Multimedia and Security Workshop (MMSEC '08), Oxford, UK, ACM (September 2008) 33–41
15. Kay, S.: Fundamentals of Statistical Signal Processing: Detection Theory. Volume 2. Prentice-Hall (1998)
16. Nikolaidis, A., Pitas, I.: Asymptotically optimal detection for additive watermarking in the DCT and DWT domains. IEEE Transactions on Image Processing **12**(5) (May 2003) 563–571
17. Tsihrintzis, G., Nikias, C.: Fast estimation of the parameters of alpha–stable impulsive interference. IEEE Transactions on Signal Processing **44**(6) (June 1996) 1492–1503