УДК 004.[056.54:932.2]

Analysis of Chi-Squared Divergence Changes by Filtering of Stego Images Formed According to UNIWARD Embedding Methods

Progonov D. O.

National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute"

E-mail: *progonov@gmail.com*

Counteraction to sensitive information leakage is topical task today. Special interest is taken on early detection of hidden (steganographic) information transferring by data transmission in communication systems. Message (stego data) embedding is provided by alteration of cover files, such as digital images, according to used steganographic algorithm. Reliable detection of formed stego images requires usage of targeted stegdetector that needs a priori information about specific distortions (signatures) of cover due to data hiding. It makes detection systems vulnerable to zero-day attack – usage by malefactors the previously unknown embedding algorithms. Therefore it is required development of universal (blind) stegdetectors that are capable to reliable revealing of stego images even in case of limited or absence information about used embedding method. Creation of blind stegdetector requires determination of cover image parameters that are sensitive to any alteration caused by message hiding. As such parameters it is proposed to use information-theoretic estimations (chi-square divergence) of pixels brightness distribution distortion due to stego data embedding. For amplification of these distortions it is used image pre-processing with median and Wiener filters. The case of adaptive messages hiding in cover images according to UNIWARD methods is considered. It is revealed that usage of chi-square divergence allows reliably detection of small alteration of cover image even in case of low cover payload (less than 10 %). Different character of chi-square divergence changes for filtered images by information hiding in spatial and JPEG domains allows determine type of used embedding domain.

Key words: steganalysis; adaptive embedding methods; UNIWARD algorithm; chi-squared divergence

DOI: 10.20535/RADAP.2019.76.72-76

Introduction

Protection of confidential information is topical task today. Special interest is taken to counteraction of information leakage by data transmission in communication systems, such as social networks [1]. Usage of steganographic methods allows hiding information into transmitted (cover) files, such as digital images. It complicates revealing of unauthorized confidential information transmission with usage of Data Leak Prevention systems (DLP-systems).

Ensuring of high detection accuracy (more than 90%) of modified (stego) images requires a priori information about specific alterations (signatures) of cover image caused by message hiding [1]. But in most real cases this information are limited or even absent, especially for advanced embedding algorithms. Therefore, it is needed development of universal (blind) stegdetectors that allows reliably detecting of stego images even in case of limitation the information about features of used embedding method.

1 Related works

One of the most widespread approaches to stego images detection is usage of covers image models [1]. These models are taking into account cover image parameters, such as [1]: parameters of pixels brightness distributions, correlation of adjacency pixels brightness, etc. Achieving of high detection accuracy requires merging several simple models into one rich model, for instance J+SRM [2], SRM [3] models. Despite of high effectiveness of mentioned cover image rich models, theirs practical applications are limited. It is caused by necessity of time-taking and computation intensive tuning of the enormous number of model's parameters (from 12,870 parameters for PSRM model to 35,263 parameters for J+SRM model) for achieving high detection accuracy.

The development of adaptive embedding methods, such as UNIWARD [4], leads to considerable reduction of models-based stegdetectors performance. It is caused by minimization of cover parameters alterations by information hiding. For increasing detection accuracy

it was proposed detection methods that are based on applying of Artificial Neural Networks (ANN) [5]. The ANN is used for extraction features from analyzed images and further construction of cover image model. Nevertheless, duration of ANN-based stegdetectors tuning remains relatively long. Therefore it is needed development of fast and accurate universal (blind) stegdetectors that are effective even in case of absence a priori information about embedding method features.

For solving of mentioned task it was proposed to analyze alterations of cover pixels brightness distortions due to information hiding with usage of information-theoretic indices, such as chi-square divergence [6]. It allows decreasing duration of stegdetector tuning by preserving of detection accuracy. Practical application of proposed solution requires usage of initial (undistorted) cover images that are hard to obtain in real cases. For this limitation to be overcome we propose to use the calibrated cover image, obtained by suppression (filtering) of specific alterations caused by message hiding.

2 Task and challenges

Our purpose is analysis of chi-square divergence changes by stego image pre-processing with usage of median and Wiener filters in case of adaptive message hiding according to UNIWARD embedding method.

3 Adaptive methods for data embedding in digital images

For minimization of cover image distortions caused by information embedding there has been proposed adaptive embedding methods (AEM). Feature of AEM is representation of data embedding process as optimization task [7]:

$$F(C,S)=\sum_{y_{\mathcal{M}}\in\mathbb{Y}}\pi(y_{\mathcal{M}})\cdot D(y_{\mathcal{M}})\to min \tag{1}$$

with limitations

$$\begin{cases}\pi(y_{\mathcal{M}})=const,\\ D(y_{\mathcal{M}})\leqslant D_{\epsilon},\end{cases} \tag{2}$$

where $F(C,S)$ – function for estimation of cover image C alterations by stego image S formation; $y_{\mathcal{M}}$ – set of cover pixels brightness changes that is needed for message $\mathcal{M}$ hiding; $\mathbb{Y}$ – set of all the possible changes of cover pixels brightness; $\pi(y_{\mathcal{M}})$ – estimation of probability a stego image formation by usage of $y_{\mathcal{M}}$; $D(y_{\mathcal{M}})$ – estimation of cover image parameters alterations due to $y_{\mathcal{M}}$ applying; D_{ϵ} – fixed level of cover parameters distortion by stego image formation.

The standard way to solving (1) is usage of assumption that embedding a single stego bit leads to independent changes of cover image parameters. In this case, functions $\pi(y_{\mathcal{M}})$ and $D(y_{\mathcal{M}})$ can be represented with usage of Gibbs distribution [7]:

$$\pi(y_{\mathcal{M}})=\prod_{i=1}^{d}\frac{exp(-\lambda\cdot D(y_i))}{\sum\limits_{y_i\in y_{\mathcal{M}}}exp(-\lambda\cdot D(y_i))},\lambda\in[0,+\infty), \tag{3}$$

$$D(y_{\mathcal{M}})=\sum_{i=1}^{d}\rho(y_i), \tag{4}$$

where d – stego bits number; $\rho(y_i)$ – function for estimation a cover image alterations by applying i-th element of set $y_{\mathcal{M}}$. Variation of Gibbs distribution parameter λ (3) allows either improving robustness of formed stego image to steganalysis ($\lambda\to+\infty$), or increasing number of embedded stego bits ($\lambda\to 0$).

One of the most robust to steganalysis embedding methods is UNIWARD groups of algorithms. These methods are based on minimization of heuristically defined distortion function $\rho(\cdot)$ [4]:

$$\rho(C,S)=\sum_{k=1}^{3}\sum_{u=1}^{N}\sum_{v=1}^{M}\frac{|W_{uv}^{(k)}(C)-W_{uv}^{(k)}(S)|}{\sigma+|W_{uv}^{(k)}(C)|}, \tag{5}$$

where C,S – grayscale cover and stego images with size $M\times N$ pixels; $W_{uv}^{(k)}(\cdot)$ – wavelet coefficient on the uv-th spatial position in the k-th sub-band of first level of two-dimensional discrete wavelet transformation; $\sigma(\sigma>0)$ – stabilizing constant.

Usage of distortion function (5) by solving of optimization task (1) with limitations (2) allows creating the uniform approach to embedding methods construction. This approach can be adapted for message embedding in spatial (S-UNIWARD algorithm) as well as JPEG (J-UNIWARD algorithm) domains.

It is shown [6] that despite of high "adaptability" of UNIWARD embedding methods to cover image, theirs usage leads to significant alteration of cover image pixel brightness distribution. These alterations can be detected by changes of information-theoretic indices, such as chi-squared divergence, by taking of initial (undistorted) cover and stego images. In most practical cases such cover image is unavailable and can be only approximated (calibrated). Therefore, it is needed analysis of information-theoretic indices changes for initial and calibrated (filtered) analyzed images.

4 Analysis of divergences between pixel brightness distributions of filtered stego images

The well-known information-theoretic indices for estimation differences between distributions of cover P_C and stego P_S images pixels brightness are [1, 8]:

Kullback-Leibler divergences (D_{KL}), Hellinger (D_H) and Bhattacharaya (D_B) distance:

$$D_{KL}(P_C, P_S) = \sum_{q \in \mathbf{Q}} P_C(q) \cdot \log_2 \left(\frac{P_C(q)}{P_S(q)} \right), \quad (6)$$

$$D_H(P_C, P_S) = \frac{1}{\sqrt{2}} \sqrt{\sum_{q \in \mathbf{Q}} \left(\sqrt{P_C(q)} - \sqrt{P_S(q)} \right)^2}, \quad (7)$$

$$D_B(P_C, P_S) = -\ln(1 - D_H^2(P_C, P_S)), \quad (8)$$

where $\mathbf{Q} = \{0, 1, \ldots, 2^k - 1\}$ – range of pixels brightness q; k – number of bits that is used for pixel's brightness encoding.

It is shown [6] that Kullback-Leibler divergence (6), Hellinger (7) and Bhattacharaya (8) distances are insensitive to weak alterations of cover pixels brightness distribution, caused by usage of AEM. For revealing these alterations it was proposed to use specific information-theoretic estimation, such as chi-squared divergence D_{χ^2} [6].

Chi-squared divergence D_{χ^2} between distributions of cover P_C and stego P_S images pixels brightness can be estimated according to formula [8]:

$$D_{\chi^2}(P_C, P_S) = \sum_{q \in \mathbf{Q}} \frac{(P_C(q) - P_S(q))^2}{P_S(q)}. \quad (9)$$

It should be noted that distance (9) is not symmetric – $D_{\chi^2}(P_C, P_S) \neq D_{\chi^2}(P_S, P_C)$. Therefore, it is useful to take into account by digital images steganalysis estimation for cover ($D^C_{\chi^2}$) and stego ($D^S_{\chi^2}$) images as well as relative ($D^{rel}_{\chi^2}$) chi-squared distances:

$$D^C_{\chi^2} = D_{\chi^2}(P_S, P_C), \quad (10)$$

$$D^S_{\chi^2} = D_{\chi^2}(P_C, P_S), \quad (11)$$

$$D^{rel}_{\chi^2} = D^C_{\chi^2} / D^S_{\chi^2}. \quad (12)$$

According to the results of [6], the relative chi-squared divergence $D^{rel}_{\chi^2}$ (12) is more sensitive to weak alterations of cover image in comparison with $D^C_{\chi^2}$ (10) and $D^S_{\chi^2}$ (11).

Typical values of $D^{rel}_{\chi^2}$ divergence by analysis of undistorted cover and stego image, formed according to S-UNIWARD and J-UNWARD embedding methods analysis are represented on Fig. 1 [6].

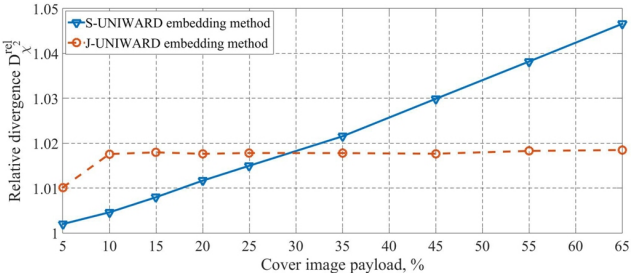


Fig. 1. Dependency of chi-squared divergence $D^{rel}_{\chi^2}$ on cover image payload for undistorted covers and stego images formed according to S-UNIWARD (solid line) and J-UNIWARD (dashed line) adaptive embedding methods.

It should be noted weak dependency of chi-squared divergence $D^{rel}_{\chi^2}$ on cover image payload in case of usage UNIWARD embedding methods (Fig. 1). Therefore, detection accuracy for standard stegdetector [9], tuned with estimated chi-squared divergence $D^{rel}_{\chi^2}$ is relatively low – probability of misclassification is $P_{MC} \approx 0.43$.

For strengthening the $D^{rel}_{\chi^2}$ changes due to message hiding it is represent an interest to pre-process analyzed image with usage of median and Wiener filters.

5 Experiments

Investigation was carried out with usage of pseudo randomly chosen 10,000 images from standard dataset MIRFLickr-25k [10]. Images were scaled to size 640x480 pixels with usage of Lanczos kernel and transformed to grayscale mode (8 bits color depth). Prepared images were saved in lossless JPEG format.

Stego images were formed according to S-UNIWARD and J-UNIWARD embedding methods [4]. Cover image payload was changed from 5% to 65% with step 5%.

Processing of images with median and Wiener filters were conducted in several steps. Firstly, analyzed image was divided into parts with usage of sliding windows W with size $w \times w (w \in \mathbb{N}_{odd})$ pixels, where $\mathbb{N}_{odd}$ is set of odd numbers. Size of sliding window was chosen $w = 7$ according to recommendation [11]. Then, for suppressing the edge effects $(w+1)/2$ points outside the image boundaries were symmetrically filled.

Image processing was started from left-upper corner and iteratively continuing by sliding window shift on 1 pixel from left to right. If sliding window achieved the last pixel in x row, window was moved to $(x+1)$ row.

For median filter, the initial (un-noised) value of central pixel brightness W^m_c was estimated as median of brightness distribution for pixels located within window.

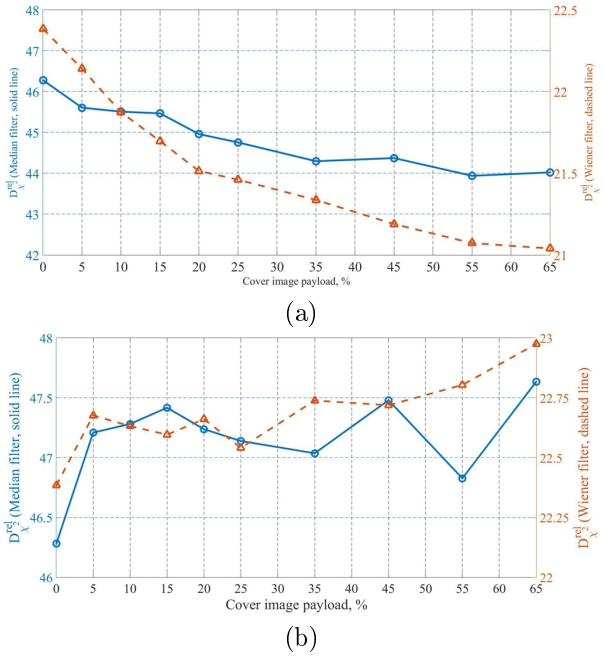
For Wiener filter, it was used standard assumption that noise is additive white Gaussian (AWGN). Estimation initial brightness of sliding window's central

pixels was calculated according to [12]:

$$W_c^w = \mu + \frac{\sigma^2 - \nu^2}{\sigma^2}\left(W_{(w+1)/2,(w+1)/2} - \mu\right),$$

where $\mu = \mathbb{E}[W]$ – expectation of pixels brightness; $\sigma^2 = (\mathbb{E}[W^2] - \mu^2)$ – variation of pixels brightness; $\mathbb{E}[\cdot]$ – expectation operator; $\nu^2 = \mathbb{E}[\sigma^2]$ – estimation of AWGN variance.

Estimation of chi-squared divergence $D_{\chi^2}^{rel}$ for filtered cover and stego images was carried out according to (12). Dependencies of divergence $D_{\chi^2}^{rel}$ on cover image payload for processed initial (undistorted) cover as well as stego images for S-UNIWARD and J-UNIWARD embedding methods are represented at Fig. 2.



(a)

(b)

Fig. 2. Dependencies of chi-squared divergence $D_{\chi^2}^{rel}$ on cover image payload for S-UNIWARD (a) and J-UNIWARD (b) embedding methods. Results for median and Wiener filters are marked as solid and dashed lines respectively.

It should be noted that different character of $D_{\chi^2}^{rel}$ changes by cover image payload variation for S-UNIWARD and J-UNIWARD embedding methods (Fig. 2). These differences can be used as distinctive features for revealing information about message embedding domain.

Also, there is revealed jump of $D_{\chi^2}^{rel}$ for stego images, even in case of low (less than 10%) cover image payload (Fig. 2). It allows using thresholding of $D_{\chi^2}^{rel}$ values for stego images detection. Application of tuned stegdetector gives opportunity to significantly decrease probability of misclassification ($P_{MC} \approx 0.17$) in comparison with case of undistorted image analysis ($P_{MC} \approx 0.43$).

Conclusion

Based on the results the performed analysis, it is shown that pre-processing (filtering) of analyzed images allows discerning differences between pixels brightness distributions of cover and stego images, formed according to advanced S-UNIWARD and J-UNIWARD embedding methods. Different character of chi-square divergence changes for stego images, formed according to S-UNIWARD and J-UNIWARD methods, can be used as distinctive features for revealing information about information embedding domain.

Taking into account these differences between chi-square divergence for cover and stego images allows up to 4 times reducing probability of misclassification (from for case of undistorted image analysis to) even in case of low cover image payload (less than 10%).

References

[1] Fridrich J. (2009) *Steganography in Digital Media*. DOI: 10.1017/cbo9781139192903

[2] Kodovský J. and Fridrich J. (2012) Steganalysis of JPEG images using rich models. *Media Watermarking, Security, and Forensics 2012*. DOI: 10.1117/12.907495

[3] Fridrich J. and Kodovsky J. (2012) Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, Vol. 7, Iss. 3, pp. 868-882. DOI: 10.1109/tifs.2012.2190402

[4] Holub V., Fridrich J. and Denemark T. (2014) Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, Vol. 2014, Iss. 1. DOI: 10.1186/1687-417x-2014-1

[5] Davidson J., Bergman C. and Bartlett E. (2005) An artificial neural network for wavelet steganalysis. *Mathematical Methods in Pattern and Image Analysis*. DOI: 10.1117/12.615280

[6] Progonov D. (2018) Information-Theoretic Estimations of Cover Distortion by Adaptive Message Embedding. *Information Theories and Applications*, Vol. 25, No 1, pp. 47-62.

[7] Filler T. and Fridrich J. (2010) Gibbs Construction in Steganography. *IEEE Transactions on Information Forensics and Security*, Vol. 5, Iss. 4, pp. 705-720. DOI: 10.1109/tifs.2010.2077629

[8] Bishop C. (2006) *Pattern Recognition and Machine Learning*, Springer-Verlag, 738 p.

[9] Kodovsky J., Fridrich J. and Holub V. (2012) Ensemble Classifiers for Steganalysis of Digital Media. *IEEE Transactions on Information Forensics and Security*, Vol. 7, Iss. 2, pp. 432-444. DOI: 10.1109/tifs.2011.2175919

[10] Huiskes M.J. and Lew M.S. (2008) The MIR flickr retrieval evaluation. *Proceeding of the 1st ACM international conference on Multimedia information retrieval - MIR '08*. DOI: 10.1145/1460096.1460104

[11] Avcibas I., Memon N. and Sankur B. (2003) Steganalysis using image quality metrics. *IEEE Transactions on Image Processing*, Vol. 12, Iss. 2, pp. 221-229. DOI: 10.1109/tip.2002.807363

[12] Gonzalez R.C and Woods R. E. (2007) *Digital Image Processing*, Prentice Hall, 976 p.

Аналіз змін хі-квадрат відстані між розподілами яскравості пікселів при фільтрації стеганограм, сформованих згідно методу UNIWARD

Прогонов Д. О.

Протидія несанкціонованій передачі конфіденційних даних є актуальною та важливою задачею сьогодні. Особлива увага приділяється ранньому виявленню прихованої (стеганографічної) передачі інформації при обміні повідомленнями в інформаційно-комунікаційних системах. Приховання повідомлень (стегоданих) проводиться шляхом внесення змін до файлів-контейнерів, зокрема цифрових зображень. Забезпечення високої імовірності виявлення сформованих стеганограм потребує застосування спеціалізованих стегодетекторів, заснованих на використанні апріорних даних щодо використаного стеганографічного алгоритму. Це призводить до зниження ефективності системи виявлення у випадку атаки нульового дня (zero-day attack) – використання зловмисниками попередньо невідомих методів прихованням повідомлень. Внаслідок цього актуальною задачею є розробка універсальних (сліпих) стегодетекторів, здатних з високою точністю виявляти стеганограми в умовах обмеженості або навіть відсутності апріорних даних щодо використаного стеганографічного алгоритму. Вирішення даної задачі потребує виявлення та аналізу слабких змін параметрів зображення-контейнеру, обумовлених вбудовуванням стегоданих. Для підсилення даних змін в роботі запропоновано проводити попередню обробку (фільтрацію) досліджуваних зображень з використанням медіанного та вінеровського фільтрів. Розглянуто випадок формування стеганограм з використанням новітніх адаптивних методів UNIWARD. Показано, що попередня фільтрація стеганограм дозволяє виявити слабкі відмінності в розподілі значень яскравості пікселів зображень-контейнерів та стеганограм, навіть у випадку малого заповнення контейнерів стегоданими (менше 10 %). Виявлено, що характер змін хі-квадрат відстані між розподілами значень яскравості пікселів зображень-контейнерів та стеганограм суттєво залежить від області вбудовування стегоданих до контейнеру. Врахування даних змін при проведенні стегоаналізу цифрових зображень дає можливість визначати область приховання повідомлень та, відповідно, обирати ефективні методи деструкції стеганограм.

Ключові слова: стегоаналіз; адаптивні стеганографічні методи; метод UNIWARD; хі-квадрат відстань

Анализ изменений хи-квадрат расстояния между распределениями яркости пикселей при фильтрации стеганограмм, сформированных согласно методу UNIWARD

Прогонов Д. А.

В работе исследованы изменения хи-квадрат расстояния между распределениями значений яркости пикселей изображений-контейнеров и стеганограмм при проведении фильтрации. Рассмотрен случай формирования стеганограмм с использованием адаптивных методов UNIWARD. Показано, что использование медианного и винеровского фильтров дает возможность выявлять слабые различия в распределениях яркости пикселей изображений-контейнеров и стеганограмм, даже в случае малого заполнения контейнеров стегоданными (менее 10 %). Выявлено, что характер изменений хи-квадрат расстояния между распределениями значений яркости пикселей исходных и обработанных стеганограмм существенно зависит от области встраивания стегоданных в изображение-контейнер. Анализ данных изменений при проведении стегоанализа цифровых изображений дает возможность определять область встраивания стегоданных в изображение-контейнер, что представляет интерес для выбора эффективных методов деструкции сформированных стеганограмм.

Ключевые слова: стегоанализ; адаптивные стеганографические методы; метод UNIWARD; хи-квадрат расстояние