Applied Soft Computing 48 (2016) 240–253

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

On digital image trustworthiness

Donghui Hu [a,*], Xiaotian Zhang [a], Yuqi Fan [a], Zhong-Qiu Zhao [a], Lina Wang [b], Xintao Wu [c], Xindong Wu [d]

[a] College of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
[b] College of Computer Science, Wuhan University, Wuhan 430072, China
[c] Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR 72701, USA
[d] Department of Computer Science, University of Vermont, Burlington, VT 05405, USA

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form 10 January 2016
Accepted 2 July 2016
Available online 16 July 2016

Keywords:
Digital image
Trustworthiness evaluation
Dempster-Shafer theory
Information fusion

ABSTRACT

Digital images are facing a crisis of trustworthiness with the emergence of various digital image processing and steganography tools. This paper proposes a novel approach that can evaluate the trustworthiness of a digital image. In this approach, an information fusion method is used to combine base digital image forensic models at the feature level and the decision level. When using different kinds of base forensic models to get supporting evidence for different kinds of digital image manipulations, there exist uncertainties introduced by base forensic models and conflicts among evidence provided by different forensic models. We use the Dempster-Shafer (D-S) evidence theory and an improved least square method to tolerate the uncertainties of forensic models and reduce the evidence conflicts. The lower and upper limits of digital image trustworthiness can then be reliably evaluated by the D-S theory. Three information fusion models based on the D-S theory are proposed. The first model uses the D-S theory at the feature fusion level. The second uses the D-S theory at the decision fusion level, where an improved least square method is designed to reduce the evidence conflicts. The last model is a combination of the first and the second one, where the D-S theory is applied at both the feature fusion and decision fusion levels. Experiments are carried out on four kinds of digital image manipulations. The experimental results show that the three proposed models are very stable in evaluating different kinds of natural images and tampering images. While the first model can only give the upper limit of the trustworthiness of a digital image, the second and the third one can give both lower and upper limits of the trustworthiness of a digital image, as well as the uncertainties of the evidence produced by base forensics models. Compared with the second model, the third one can further reduce the uncertainties. The experimental results also show that when a digital image undergoes many kinds of manipulations, our models can validly compute a soft degree to measure the trustworthiness of the image, while current ordinary digital image forensic models may fail to predict it correctly. Experimental results also demonstrate that the proposed digital image trustworthiness evaluation models can be adapted as digital image forensic classification models with very high detection accuracy.

1. Introduction

As the representation of natural scenes, digital images traditionally imply the truth and fact. However, with the popularization of various digital image processing tools, people can easily manipulate digital images. Digital images can also be used to hide secret information by various steganography tools. Those manipulations greatly undermine the credibility of digital images. Nowadays, people are often concerned about whether a digital image is trustworthy before they are comfortable to use it. As a result, digital images are facing crisis of trustworthiness.

Trust is a common phenomenon of our human life. It has been argued that we as humans would not even be able to face the complexities of the world without resorting to trust [1]. Many literatures have modeled trust from different perspectives. The Oxford Reference Dictionary states the trust as the firm belief in the reliability or truth or strength of an entity. Golembiewski et al. [2] think that trust "implies some degree of uncertainty as to outcome", and trust also "implies hopefulness or optimism as to outcome".

* Corresponding author.
E-mail addresses: hudh@hfut.edu.cn (D. Hu), msadsl6233925@126.com (X. Zhang), yuqi.fan@gmail.com (Y. Fan), zhongqiuzhao@gmail.com (Z.-Q. Zhao), lnawang@163.com (L. Wang), xintaowu@uark.edu (X. Wu), xwu@cs.uvm.edu (X. Wu).

http://dx.doi.org/10.1016/j.asoc.2016.07.010

Yamamoto et al. [3] state that “the decision to trust is based on evidence to believe, or be confident in, someone’s or something’s good intentions towards us”. Grandison et al. [4] consider that trust is a composition of many different attributes – reliability, dependability, honesty, truthfulness, security, competence, and timeliness, which may have to be considered depending on the environment in which trust is being specified.

In this paper, we consider the scenario that, when people or software use a digital image, they may be concerned about its integrity, origin and/or security. People tend to use more trustworthy images than un-trustworthy ones. Inspired by the trust definition of Grandison et al. [4], in this paper, we propose models to measure the trustworthiness of a digital image using different trustworthiness attributes. We term our models as digital image trustworthiness evaluation models. The trustworthiness of a digital image is a value that tells how trustworthy the image can be used. Digital images may have many trustworthiness attributes or trustworthiness evaluation indexes, such as the index of security (to evaluate whether the image contains hidden secret information), the index of integrity (to evaluate whether the digital image is manipulated), the index of authenticity (to evaluate whether the image is generated by a digital camera or by a computer graphics render tool), and so on. Different digital image users may have concerns with different trustworthiness attributes. In this paper, we propose models to effectively calculate the degree of a digital image’s trustworthiness.

Many techniques associated with trustworthiness of digital images have been developed in recent years. Active techniques such as digital watermarking [5] and perceptual image hash [6,7] can ensure the trustworthiness of a digital image. In these cases, the original image must be in hand, and the corresponding digital watermarking should be embedded (or perceptual image hash should be generated). However, in real life, most of the digital images do not contain digital watermarking nor have an image hash. Digital image forensics [8] is a passive technique which can determine whether an image is tampered or not. However, it cannot capture the degree of trustworthiness for a given image, nor can measure trustworthy attributes (e.g. the origin, integrity and security) synthetically. In this paper, an effective method that can measure the trustworthiness of a digital image synthetically and quantitatively will be developed.

Moreover, in this paper, when designing methods to evaluate the trustworthiness of a digital image, we use various digital image forensic models to capture the image’s different trustworthiness attributes. We find that the evidence supported by different forensic models may conflict with each other, and each forensic model comes with uncertainties when predicting multi-manipulations [9]. To achieve more reliable and effective digital image trustworthiness evaluation, we propose the use of the Dempster-Shafer (D-S) evidence theory and data fusion. The D-S theory [10,11], which is generally used to combine separate pieces of evidence to calculate the probability of an event, is a mathematical theory of evidence based on belief functions and plausible reasoning. The D-S evidence theory allows the representation of ignorance by giving support to the case of “unknown”. It can reduce the uncertainty by allowing evidence to support several mutually exclusive conclusions, and can represent data imperfections without the need to make simplifying assumptions [12]. Data fusion is a method to combine data from multiple sensors in order to achieve better accuracy and more specific inferences than those using a single sensor alone [13]. It can be applied at the raw data level, feature level, or decision level [13,14].

Three fusion models are presented based on the D-S theory and data fusion. The first uses feature fusion to address the problem of evidence conflict when using Dempster’s rule of combination. The second uses the Dempster’s rule of combination at the decision fusion level, in which we develop an improved least square method to deal with the evidence matrix and reduce the conflicts among evidence provided by different forensic models. Uncertainty of each forensic model is calculated in the training phase and then combined in the evaluation phase. The lower and upper limits of digital image trustworthiness are calculated by using the Dempster’s rule of combination. The last model is a combination of the first and second ones where the uncertainty is further reduced by the feature fusion.

Finally, we conduct simulations on 11,200 test images. Simulation results show that the proposed three models can evaluate digital image trustworthiness effectively and reliably, and the last one performs the best.

The rest of this paper is organized as follows. Section 2 reviews related work and compares the trustworthiness evaluation methods with current digital forensic ones. Section 3 describes three models based on the D-S theory and information fusion. The effectiveness of the three models is evaluated and compared in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

Digital image trustworthiness evaluation models are interactively related to digital image forensic models, as shown in Fig. 1. In our proposed digital image trustworthiness evaluation models, we use digital image forensic models to capture different aspects of tampering features and provide forensic decision or evidence for each kind of tampering. The digital image trustworthiness

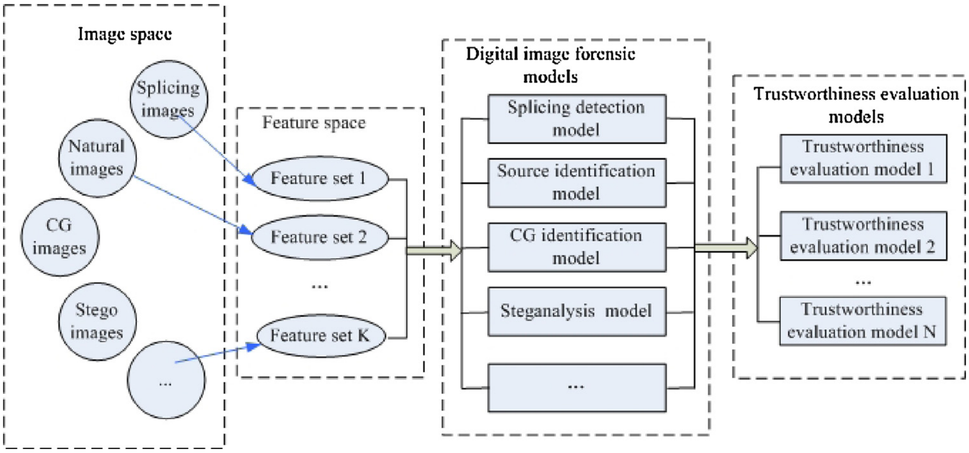


Image space
Splicing images
Natural images
CG images
Stego images
...
Feature space
Feature set 1
Feature set 2
...
Feature set K
Digital image forensic models
Splicing detection model
Source identification model
CG identification model
Steganalysis model
...
Trustworthiness evaluation models
Trustworthiness evaluation model 1
Trustworthiness evaluation model 2
...
Trustworthiness evaluation model N

Fig. 1. Relationship between digital image trustworthiness evaluation models and digital image forensic models.

evaluation models combine the decision results (or tampering evidence) of different forensic models to compute a soft score which measures the image's trustworthiness. So in this section, we first discuss the related research work in digital image forensics, and then describe why we propose digital image trustworthiness evaluation models based on forensic ones.

Digital image forensics can be classified into four categories:

- Image forgery detection which is to determine whether an image has undergone any form of manipulations after it was initially captured [8]. Numerous methods for image forgery detection have been reported, most of which are based on local inconsistencies, such as lighting inconsistencies [15,16], camera parameter or statistic inconsistencies [17,18], and high order statistic inconsistencies [19,20]. Additionally, JPEG image re-compression detection is investigated [21,22] since it is important to accurately blind JPEG image steganalysis [23]. Recently, several techniques (such as counter-forensics [24,25]) have been developed to erase or camouflage the traces left by the applications of processing operations, which make the detection of image forgery more difficult. As a result, many researchers have designed new forensic methods such as forensics of image resizing [26,27], contrast enhancement [28,25] and median filtering [29,30].
- Image source identification which is to determine where or how an image is generated. For example, source camera identification can be divided into two categories. The first one determines whether an image is generated from the same type of camera, based on common image statistics (such as CFA and demosaicing artifacts) made by the same camera brand [31,32]. The other one determines whether an image is generated from a particular camera, based on individual statistics results from sensor dust or sensor imperfection [33,34].
- Computer generated (CG) image identification. As a result of computer vision and computer simulation, photo-realistic computer generated images become very similar to natural images. Some approaches have been proposed to detect CG images from natural images [35,36]. For example, a statistical model for photographic images consisting of first- and higher-order wavelet statistics was proposed in [35]. A scheme based on statistical moments of a wavelet sub-band's histogram in DFT domain was proposed in [36].
- Forensic steganalysis. Some researches take steganalysis as a kind of forensics, and call it forensic steganalysis [37,38]. How to assure that the result of steganalysis can be submitted as legal evidence is still a challenging issue. Studies on recovering the stego key or extracting the hidden information [38,39] may contribute to the solution of the problem.

Current research mainly focuses on a single tampering, while in the real life, digital images may undergo more than one tampering. When using several unreliable digital image forensics tools (models) simultaneously to predict multi-tampering of a digital image, uncertainty is often introduced by these different forensics tools (models) [9,40]. Barni et al. [40] proposed a fuzzy logic method to deal with this kind of "uncertainty" in image forensics, but the method can neither measure the uncertainty nor tell the degree of the trustworthiness of the given digital image. Zhang et al. [41] conducted a survey that inconsistencies exist in digital image forensics, such as the lighting inconsistency [42], the local noise level inconsistency [43], and the blocking artifact inconsistencies. When using several digital image forensic models simultaneously to predict a multi-tampering image, inconsistencies or conflicts may occur among the evidence produced from different forensic models. This is because one evidence may support a kind of tampering while another evidence may support another kind of tampering. We call this phenomenon evidence conflict. For each forensic model, there may be uncertainty because in some cases a forensic model cannot predict the real tampering class of the image accurately especially when the digital image undergoes many kinds of tampering. We call this phenomenon uncertainty of the forensic model (the formal definition of uncertainty will be given in Section 3). We develop digital image trustworthiness evaluation models (the second and the third one) based on the D-S theory which can compute the upper and lower limits of the trustworthiness of a digital image. The uncertainty of trustworthiness is the difference between the upper limit and lower limit of the trustworthiness, so in this sense we say our model can allow for uncertainty of the forensic model. Moreover, our second and third models, we use an improved least square method to reduce evidence conflict before we use these evidence to compute the final trustworthiness of a digital image. Our approach can accurately evaluate the degree of the trustworthiness of a digital image, which can be used as a reference value in digital image retrieval [44] or digital image fusion methods. Our proposed method, to the best of our knowledge, is the first one that intends to compute (evaluate) the digital image trustworthiness, and at the same time resolve the problem of uncertainty and inconsistency.

Current research mainly focuses on a single tampering, while in the real life, digital images may undergo more than one tampering. When using several unreliable digital image forensics tools (models) simultaneously to predict multi-tampering of a digital image, uncertainty is often introduced by these different forensics tools (models) [9,40]. Barni et al. [40] proposed a fuzzy logic method to deal with this kind of "uncertainty" in image forensics, but the method can neither measure the uncertainty nor tell the degree of the trustworthiness of the given digital image. Zhang et al. [41] conducted a survey that inconsistencies exist in digital image forensics, such as the lighting inconsistency [42], the local noise level inconsistency [43], and the blocking artifact inconsistencies. When using several digital image forensic models simultaneously to predict a multi-tampering image, inconsistencies or conflicts may occur among the evidence produced from different forensic models. This is because one evidence may support a kind of tampering while another evidence may support another kind of tampering. We call this phenomenon evidence conflict. For each forensic model, there may be uncertainty because in some cases a forensic model cannot predict the real tampering class of the image accurately especially when the digital image undergoes many kinds of tampering. We call this phenomenon uncertainty of the forensic model (the formal definition of uncertainty will be given in Section 3). We develop digital image trustworthiness evaluation models (the second and the third one) based on the D-S theory which can compute the upper and lower limits of the trustworthiness of a digital image. The uncertainty of trustworthiness is the difference between the upper limit and lower limit of the trustworthiness, so in this sense we say our model can allow for uncertainty of the forensic model. Moreover, our second and third models, we use an improved least square method to reduce evidence conflict before we use evidence to compute the final trustworthiness of a digital image. Our approach can accurately evaluate the degree of the trustworthiness of a digital image, which can be used as a reference value in digital image retrieval [44] or digital image fusion methods. Our proposed method, to the best of our knowledge, is the first one to compute (evaluate) the digital image trustworthiness, and at the same time resolve the problem of uncertainty and inconsistency.

3. Digital image trustworthiness evaluation models

As mentioned in Section 2, when using multiple forensic models simultaneously to predict multi-tampering of a digital image, there exist uncertainty and inconsistency problems, because digital images may undergo many kinds of tampering or processes in the real life applications. In this section, we propose new models based on the D-S theory and data fusion to evaluate the trustworthiness of digital images. Our models can effectively resolve uncertainty resulting from different forensic models and conflicts among different evidence.

3.1. Problem formulation and primitives

We first introduce the D-S theory. In the D-S theory, parameters such as events in probability theory, are called propositions.

Definition 1. The set of finite parameters $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is called the frame of discernment, where θ_i is a proposition.

Unlike the traditional probability theory, in the D-S theory of belief-function, the evidence (or uncertainty) is not only assigned to the single element of the frame, but also assigned to all other proper subsets of the frame and to the entire frame Θ.

Definition 2. Set Θ as a frame of discernment. The function $m{:}2^{\Theta} \to [0, 1]$ is called a basic belief assignment function if

$$\sum_{A_i \subset \Theta} m(A_i) = 1 \tag{1}$$

and

$$m(\emptyset) = 0 \tag{2}$$

where each A_i supporting $m(A_i) > 0$ is called a focal set, and $m(A_i)$ represents the degree of the evidence supporting the proposition of A_i.

The basic belief assignment function is also called mass function. In our proposed digital trustworthiness evaluation model, we use base forensic models to detect tampering and the base forensic models return probabilities as mass functions. Those probabilities are used as supporting evidence for different kinds of tampering.

Definition 3. Belief function $Bel : 2^{\Theta} \to [0, 1]$

$$Bel(A) = \sum_{A \supseteq B} m(B) \tag{3}$$

where B is a subset of A.

$Bel(A)$ can be regarded as the degree of belief that A is true. Notice that the quantity $m(A)$ measures the belief that one commits exactly to A, not the total belief that one commits to A. Quantity $m(B)$ for all subsets B of A must be added to measure the total belief function of (A).

Definition 4. The plausibility of a proposition A is the sum of all the masses of the sets that intersect with A (which means we should consider all of the possible evidence that may be involved in supporting A). So we can define the Plausibility function $Pl : 2^{\Theta} \to [0, 1]$ as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) = 1 - \sum_{\neg A \supseteq B} m(B) \tag{4}$$

also,

$$Pl(A) = 1 - Bel(\neg A) \tag{5}$$

$Bel(A)$ is the lower bound of probability of focal set A, while $Pl(A)$ is the upper bound. The precise probability of an event lies within the lower and upper bounds of belief and plausibility, respectively. The probability is uniquely determined if $Pl(A) = Bel(A)$.

The measures of $Bel(A)$ and $Pl(A)$ are derived from the combined basic assignments. The Dempster's rule combines multiple belief functions through their probability assignments. These belief functions are defined on the same frame of discernment, but are based on independent arguments or evidence.

Definition 5. Let $m_1, m_2, \cdots, m_n$ be mass functions on the same frame of Θ, then the Dempster's rule of combination is:

$$m(A) = (1 - k)^{-1} \cdot \sum_{\cap_i A_j = A} \prod_{1 \le i \le n} m_i(A_j) \tag{6}$$

where

$$k = \sum_{\cap_i A_j = \emptyset} \prod_{1 \le i \le n} m_i(A_j) \tag{7}$$

k is the conflict factor among evidences of $m_i(A_j)$.

To evaluate the trustworthiness of digital images, first of all, we should set up the trustworthiness attributes.

$T = \{T_1, T_2, \ldots, T_K\}$ are a set of trustworthiness attributes (or trustworthiness evaluation indexes), $K = \| T \|$, where each $T_i (i = 1, 2, \ldots, K)$ is a trustworthiness attribute used as an evaluation index. For example, $T = \{stego, splicing, re-compressed, CG\}$ means to evaluate the trustworthiness of an image by deciding whether it is a stego image, a splicing image, a re-compressed image, and (or) a computer generated image synthetically. Different users may apply different trustworthiness evaluation indexes in different situations. For example, if a user is a newsperson and more concerns about the origin and integrity attributes than the security attribute, she can use $T = \{splicing, re-compressed, CG\}$ as her trustworthiness evaluation indexes. In the real life applications, the system can provide digital image trustworthiness evaluation attributes as many as possible, which current forensic models can capture. The users can select the subset of the full trustworthiness evaluation attributes provided by the system according to their usage. The trustworthiness evaluation indexes also can be hierarchy (multilevel), for example, the evaluation of *stego* can have sub evaluation indexes of *MB*1 *Stego*, *MB*2 *Stego*, *Jsteg* and so on, which means we need to evaluate whether the image has embedded secret information by steganography tools of *MB*1 *Stego*, *MB*2 *Stego* and so on. In this paper, we consider one level evaluation indexes.

Definition 6. The D-S theory based digital image trustworthiness evaluation: $Trustworthiness : I \times T \to [\underline{Trust}, \overline{Trust}]$, where $\underline{Trust}$ is the lower limit of the trustworthiness of image I on evaluation indexes of T, $\overline{Trust}$ is the upper limit of the trustworthiness, and the uncertainty of trustworthiness is:

$$d = \overline{Trust} - \underline{Trust} \tag{8}$$

In our models, we use $Bel(A)$ to denote $\underline{\mathrm{Trust}}$, and use $Pl(A)$ to denote $\overline{Trust}$, so the uncertainty d is $Pl(A) - Bel(A)$.

In our simulations, we set the trustworthiness within the interval of [0,1]. If the values of $\underline{Trust}$ and $\overline{Trust}$ are near 1, which indicates the evaluated image is trustworthy. If the values of $\underline{Trust}$ and $\overline{Trust}$ are near 0, the evaluated image is not trustworthy. The higher the value d, the more uncertainty the evaluation method.

Finally we define the combination of feature sets, which will be used to train our models.

Definition 7. Combination of feature sets Com_Sets: Let $X_1, X_2, \ldots, X_K$ denote different feature sets which capture different characteristics of trustworthy attributes, for example, $X_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ captures the trustworthy attribute T_i. $Com\_Sets(X_1, X_2, \ldots, X_K)$

denotes all the combinations of $X_1, X_2, \ldots, X_K$ as:

$$\begin{aligned} &Com\_Sets(X_1, X_2, \ldots, X_K) \\ &\quad = \{X_1, X_2, \ldots, X_K\} \cup \{X_1 \mid X_2, X_1 \mid X_3, \ldots, X_{K-1} \mid X_K\} \cup \ldots \\ &\quad\ \cup \{X_1 \mid X_2 \mid \ldots \mid X_K\} \end{aligned} \tag{9}$$

where the number of possible combinations of feature sets is $2^K - 1$, and the symbol of "|" represents the operation of feature combination.

Three models are proposed in the following subsections. The first model, $DSTM_1$, is based on the feature fusion level. The second one, $DSTM_2$, is based on the decision fusion level. The third one, $DSTM_3$, is the combination of $DSTM_1$ and $DSTM_2$. The D-S theory is used in all of the three models. In $DSTM_2$ and $DSTM_3$, the improved least square method is used to pre-process the evidence matrix and reduce evidence conflicts.

3.2. DSTM$_1$

Given an image I, the problem of trustworthiness evaluation can be taken as the measurement of the proposition $A : \{I \text{ is trustworthy}\}$, which means using the user selected trustworthiness evaluation indexes to evaluate whether the image is trustworthy or not. According to the trustworthiness evaluation indexes, the proposition can be divided into $A = \{A_1, A_2, \ldots, A_K\}$. According to (5), we get

$$\begin{aligned} Pl(A) &= 1 - Bel(\neg A) = 1 - Bel(\neg\{A_1, A_2, \cdots, A_K\}) \\ &= 1 - Bel(\{\neg A_1, \neg A_2, \cdots, \neg A_K\}) \end{aligned} \tag{10}$$

According to (3), we get

$$\begin{aligned} Bel(\{\neg A_1, \neg A_2, \cdots, \neg A_K\}) = &\sum_{1 \le i \le K} m(\neg A_i) + \sum_{1 \le i, j \le, i \ne j} m(\neg A_i, \neg A_j) \\ &+ \cdots + m(\neg A_1, \neg A_2, \ldots, \neg A_K) \end{aligned} \tag{11}$$

Suppose we use a multi-classifier $h : X \times Y \rightarrow [0, 1]$ in the forensic model, where $X \in Com\_Sets(X_1, X_2, \ldots, X_K)$ is a combination of feature sets, $|X| = N$. The training data are $\{([x_{11}, x_{12}, \ldots, x_{1N}], y_1), \ldots, ([x_{M1}, x_{M2}, \ldots, x_{MN}], y_M)\}$, M is the size of the training sample, and $y_i \in Y$ are labels for multi-class (for example $Y = [0, 1, 2, \ldots, L]$). Denote $I(\pi) = 1$ if π is true, otherwise $I(\pi) = 0$. The whole model includes two phases.

Phase 1: Forensic model training

The training phase can be described as follows [45,46]:

1 initialize the training data with weight $\omega_i = \frac{1}{K}, i = 1, 2, \ldots, L$
2 for $s = 1$ to S:
  (a) for $n = 1$ to N, train a weak classifier f_n using training data of $\{(x_{1n}, y_1), \ldots, (x_{Mn}, Y_M)\}$ with ω_i.
  (b) get a middle final classifier h_s by combining $\{f_n\}$ using weighted voting.
  (c) compute $\varepsilon_s = \frac{\sum_{i=1}^{N} \omega_i (y_i \ne h_s(x_i))}{\sum_{i=1}^{N} \omega_i}$.
  (d) compute $\alpha_s = \log \frac{1 - \varepsilon_s}{\varepsilon_s}$.
  (e) set $\omega_i \leftarrow \omega_i \cdot \exp(\alpha_t \cdot I(y_i \ne h_t(x_i)))$, $i = 1, 2, \ldots, L$.
  (f) re-normalize ω_i.
3 output $h(x) = \underset{k}{argmax} \sum_{s=1}^{S} \alpha_s \cdot (h_s(x) = k)$

According to [46], each base classifier is trained on a feature set. In this model, we have $2^K - 1$ feature sets in the $Com\_Sets(X_1, X_2, \cdots, X_K)$, so in this training algorithm, $S = 2^K - 1$.

Phase 2: Trustworthiness evaluation

1 For a given image, feed the testing results of forensic classifiers to the Bradley–Terry model [47] with "one-against-the rest" setting, and get a multi-class probability of P, where $P \leftarrow [p_0, p_1, p_2, \ldots, p_K]$, $p_i = p(y_i \mid x) \rightarrow [0, 1]$, $x \in X$, $y_i \in Y$ is the probability of classifier resulting from a given test. Let $m(X) \leftarrow \sum_{y_i \in class\ \ of\ \ tampering} p_i$.
2 For each $X' \in Com\_Sets(X_1, X_2, \ldots, X_K)$, calculate $m(X')$, and calculate the average value of all the $m(X')$, then we have $Bel(\neg A) = (2^K - 1)^{-1} \cdot (\sum_{X' \in Com\_Sets(X_1, X_2, \ldots, X_K)} m(X'))$.
3 According to (10), the upper limit of the trustworthiness is $\overline{Trust} = 1 - Bel(\neg A)$.

In $DSTM_1$, we do not use the Dempster's rule of combination. Instead, we use feature combination on different possible combinations of feature sets (see (10)) to tolerate the conflicts among different evidence. We use each tampering feature and their possible combinations to train forensic models. The mass functions that support each tampering and their possible combinations are calculated in the trustworthiness evaluation phase. All possible tampering and their tampering combinations are considered to calculate $Bel(\neg A)$, and the final upper limit of the trustworthiness is $1 - Bel(\neg A)$, so in this sense, we say that the model can tolerate the conflicts among different evidence. On the contrary, because of this kind of tolerance, the model can only compute the upper limit of a digital image's trustworthiness. In the next subsection, we propose the second model, $DSTM_2$, which can reduce the conflicts among different evidence, and at the same time can calculate the uncertainties produced by different base forensic models, so that both the upper limit and the lower limit of a digital image's trustworthiness can be computed.

3.3. DSTM$_2$

In $DSTM_1$, the combination of $m(X)$ is achieved by calculating the result of a feature fusion based multi-classifier, and each $m(X)$ corresponds to one combination of the feature set. In $DSTM_2$, the uncertainties of base forensic models are calculated at the end of the training phase. The evidence matrix is pre-processed by an improved least square method to reduce the conflicts among different evidence produced by different base forensic models. The D-S theory is used at the decision fusion level, and both the upper limit and the lower limit of a digital image's trustworthiness are calculated. The whole model includes three phases as shown in Fig. 2.

Phase 1: Forensic Model Training

Suppose the size of training data is M. Let $X_1, X_2, \ldots, X_K$ be different feature sets for each forensic model, and each model produces L probabilities for L classes. The forensic probabilities for the model with the feature set X_i can be denoted as:

$$P(X_i) = \begin{bmatrix} p_{1,0}(X_i) & p_{1,1}(X_i) & \cdots & p_{1,L}(X_i) \\ p_{2,0}(X_i) & p_{2,1}(X_i) & \cdots & p_{2,L}(X_i) \\ \cdots & \cdots & \ddots & \cdots \\ p_{M,0}(X_i) & p_{M,1}(X_i) & \cdots & p_{M,L}(X_i) \end{bmatrix} \tag{12}$$

where $p_{j,k}(X_i)$ denotes the probability of image j is classified as class k in the forensic model with feature set X_i.

Definition 8. Given a threshold ε, for each $p_{j,k}(X_i)$ in matrix of $P(X_i)$, if $| p_{j,k_1}(X_i) - p_{j,k_2}(X_i) | \le \varepsilon$ $(k_1 \le k_2)$, then we say an uncertain event happens. Let $U(X_i)$ denote all the uncertain events in the whole

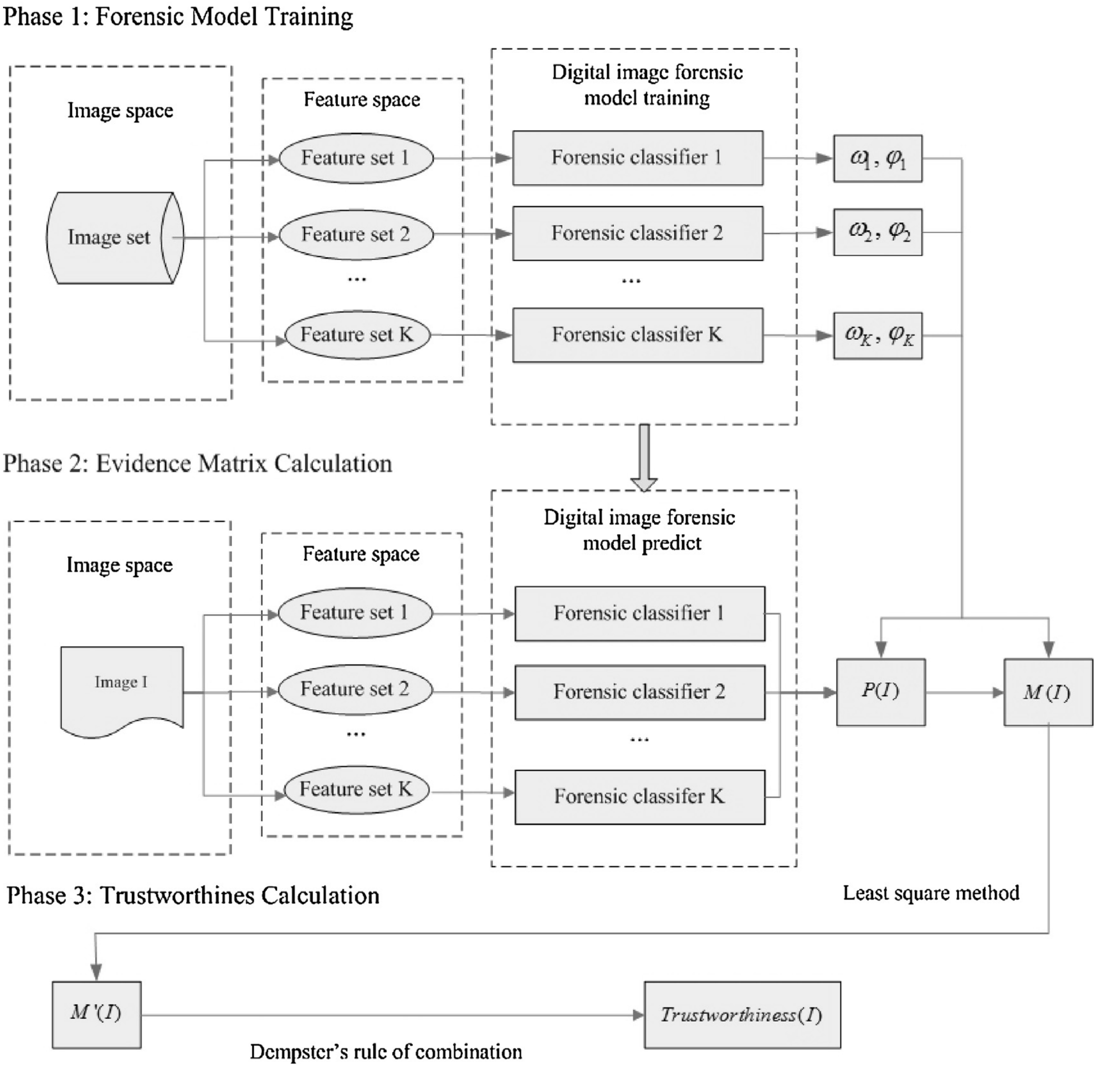


Fig. 2. *DSTM*$_2$: Applying D-S evidence theory at decision fusion level to evaluate digital image trustworthiness

training data set, then the uncertainty rate is $\varphi_i = \frac{U(X_i)}{M}$, where M is the size of training data set. The threshold is related to the applications of digital image trustworthiness evaluation. Different applications have different uncertainty requirements of trustworthiness evaluation, and may also have different trustworthiness evaluation indexes.

We compute the detection accuracy of each digital image forensic model with the feature set of X_i. We denote the normalized result as ω_i which will be used as a weight in Phase 2.

Phase 2: Evidence matrix calculation

For a given image I, let $P(I)$ denote all the probabilities of image I in all forensic models. Suppose the index of I in the total M images is j, then:

$$P(I) = \begin{bmatrix} p_{j,0}(X_1) & p_{j,1}(X_1) & \cdots & p_{j,L}(X_1) \\ p_{j,0}(X_2) & p_{j,1}(X_2) & \cdots & p_{j,L}(X_2) \\ \dots & \dots & \ddots & \dots \\ p_{j,0}(X_K) & p_{j,1}(X_K) & \cdots & p_{j,L}(X_K) \end{bmatrix} \tag{13}$$

$P(I)$ is modified as in (14) using the uncertainty rate φ_i computed from Phase 1, where the last column is a new column which denotes the uncertainty of each forensic model.

$$P'(I) = \begin{bmatrix} \frac{p_{j,0}(X_1)}{1+\varphi_1} & \frac{p_{j,0}(X_1)}{1+\varphi_1} & \cdots & \frac{p_{j,0}(X_1)}{1+\varphi_1} & \frac{\varphi_1}{1+\varphi_1} \\ \frac{p_{j,0}(X_2)}{1+\varphi_2} & \frac{p_{j,0}(X_2)}{1+\varphi_2} & \cdots & \frac{p_{j,0}(X_2)}{1+\varphi_2} & \frac{\varphi_2}{1+\varphi_2} \\ \dots & \dots & \ddots & \dots & \dots \\ \frac{p_{j,0}(X_K)}{1+\varphi_K} & \frac{p_{j,0}(X_K)}{1+\varphi_K} & \cdots & \frac{p_{j,0}(X_K)}{1+\varphi_K} & \frac{\varphi_K}{1+\varphi_K} \end{bmatrix} \tag{14}$$

We use $P'(I)$ to denote the evidence matrix, and $m_{i,k}$ to denote $\frac{p_{j,k}(X_i)}{1+\varphi_i}$. Then the evidence matrix is:

$$M(I) = \begin{bmatrix} m_{1,0} & m_{1,1} & \cdots & m_{1,L} & m_{1,L+1} \\ m_{2,0} & m_{2,1} & \cdots & m_{2,L} & m_{2,L+1} \\ \dots & \dots & \ddots & \dots & \dots \\ m_{K,0} & m_{K,1} & \cdots & m_{K,L} & m_{K,L+1} \end{bmatrix} \tag{15}$$

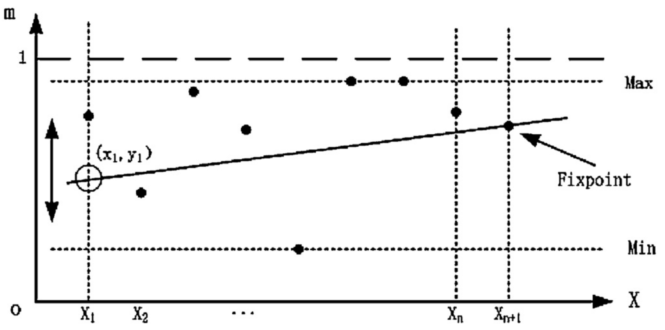


Fig. 3. Reduce the evidence conflicts using the least square method.

In $M(I)$, each row stands for a piece of evidence provided by a base forensic model, and each column stands for evidence from different base forensic models supporting a kind of tampering (except that the first column supports a natural image, and the last column supports the uncertainty of a base forensic model). For example, $m_{i,j}(I)$ $(1 \le i \le K,\ 1 \le j \le L)$ is evidence provided from the i_{th} base forensic model that supports the j_{th} kind of tampering. From experimental results we have observed that the evidence provided by different base forensic models may conflict with each other, there exist "outliers" in a column of the matrix that supports one kind of tampering. As shown in Fig. 3, some values of evidence may be very high and some others may be very low. In our method, we think each evidence has its contribution in calculating the final trustworthiness of an image, so instead of simply deleting or neglecting these "outliers", we use an improved least square method to preprocess the evidence to make the evidence with very high value or very low value "move" to the best-fit line in the center (see the line l in the center of Fig. 3) which is calculated by the improved least square method. By using the improved least square method, we can take all evidence provided by every base forensic model into account, and at the same time reduce the conflict among different evidence.

The designed improved least square method, which preprocesses an evidence matrix $M(I)$ into an improved evidence of $M'(I)$, includes four steps as follows.

1 Let $M_j[I] = [m_{1,j}, m_{2,j}, \ldots, m_{K,j}]^T$ denote a column in $M(I)$. Then we can construct K data points (x_i, y_i) with $x_i = i \cdot step$, $y_i = m_{i,j}$, where $i = 1, 2, \ldots, K$, $step$ is a constant, and $m_{i,j} \in M_j[I]$.

2 Let $x_{K+1} = (K+1) \cdot step$, and $y_{K+1} = \frac{\sum_{1 \le i \le K} \omega_i \cdot m_{i,j}}{\sum_{1 \le i \le K} \omega_i} = \sum_{1 \le i \le K} \omega_i \cdot m_{i,j}$, where ω_i is a weight which is computed from phase 2. Then we construct a new point (x_{K+1}, y_{K+1}), and use it as a fixpoint.

3 Find the best fit line $l : y = ax + b$ under the following conditions:
(a) The line l must pass through the fixpoint (x_{K+1}, y_{K+1}).
(b) Let $y_{\max} = \max\limits_{i=1,\cdots,K}(y_i)$,$y_{\min} = \min\limits_{i=1,\cdots,K}(y_i)$. As shown in Fig. 3, when $x \in [x_1, x_{K+1}]$, the line l must be between line $y = y_{max}$ and line $y = y_{min}$.
(c) The least weighted sum of squares of distance is between the K points (x_i, y_i) and the line l.

From condition a), we get:

$$y_{K+1} = ax_{K+1} + b \tag{16}$$

From condition b), we learn $ax_1 + b \ge y_{min}$ and $ax_1 + b \le y_x ax$. From (16), we get

$$a \le \frac{y_{K+1} - y_{\min}}{x_{K+1} - x_1} \tag{17}$$

and

$$a \ge \frac{y_{K+1} - y_{\max}}{x_{K+1} - x_1} \tag{18}$$

For condition c), let $S(a)$ denote the weighted sum of squares of distance. $S(a) = \sum_{i=1}^{k} \omega_i (y_i - ax_i - b)^2$. From (16), we get:

$$S(a) = \sum_{i=1}^{K} \omega_i [(y_i - y_{K+1}) + a(x_{K+1} - x_i)]^2 \tag{19}$$

$$\frac{dS(a)}{da} = 2\sum_{i=1}^{K} \omega_i (x_{K+1} - x_i)[(y_i - y_{K+1}) + a(x_{K+1} - x_i)] \tag{20}$$

$$\frac{d^2S(a)}{da^2} = 2\sum_{i=1}^{K} \omega_i (x_{K+1} - x_i) \tag{21}$$

From (21) we know $\frac{d^2S(a)}{da^2} > 0$, so $\frac{dS(a)}{da}$ is monotonically increasing. Let $\frac{dS(a)}{da} = 0$. We get

$$a_0 = \frac{\sum_{i=1}^{K} \omega_i (x_{K+1} - x_i)(y_{K+1} - y_i)}{\sum_{i=1}^{k} \omega_i (x_{K+1} - x_i)^2} \tag{22}$$

From (17) and (18), let $a_1 = \frac{y_{K+1} - y_{\max}}{x_{K+1} - x_1}$ and $a_2 = \frac{y_{K+1} - y_{\min}}{x_{K+1} - x_1}$. Then we compute the minimum of $S(a)$ from $a_1 \le a \le a_2$ as follows:
(a) if $a_0 \le a_1$, then $a = a_1$.
(b) if $a_1 \le a_0 \le a_2$, then $a = a_0$.
(c) if $a_0 \ge a_2$, then $a = a_2$.
From (16), $b = y_{K+1} - ax_{K+1}$, we get $l : y = ax + b$.

4 For each $j \in [0, L]$, modify $M_j[I]$ with $m'_{i,j} = ax_i + b(i = 1, 2, \cdots, K)$, and then

$$M'(I) = \begin{bmatrix} m'_{1,0} & m'_{1,1} & \cdots & m'_{1,L} & m'_{1,L+1} \\ m'_{2,0} & m'_{2,1} & \cdots & m'_{2,L} & m'_{2,L+1} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ m'_{K,0} & m'_{K,1} & \cdots & m'_{K,L} & m'_{K,L+1} \end{bmatrix} \tag{23}$$

It is easy to prove that the result of $M'(I)$ has no relationship with the constant $step$.

Phase 3: Trustworthiness evaluation

Normalize each row in $M'(I)$. Each column in $M'(I)$ indicates the evidence of image I belonging to type of $j, j = 0, 1, \ldots, L$. For example, $j = 0$ indicates I is a natural image and $j = 1$ denotes I is a stego image.

Then we use the Dempster's rule of combination as defined in (6) to combine the evidence.

Assuming propositions $A : \{I \, is \, trustworthy\}$, $A_0 : \{I \, is \, a \, natural \, image\}$, $A_i : \{I \, is \, tampering_i\}$, where $i = 1, 2, \cdots, L$, and $tampering_i$ is a kind of tampering, $A_{L+1} : \{I \, is \, uncertainty\}$, then

$$m'(A_0) = \frac{\sum \bigcap_{i \in [0,K], j \in [0,L+1]} {}^{A_j = A_0} \prod\limits_{i \in [0,K], j \in [0,L+1]} m_i{}'(A_j)}{(1 - K)} \tag{24}$$

where

$$k = \sum_{\cap_i A_j = \Phi} \prod_{1 \le i \le n} m'_i(A_j) \tag{25}$$

According to (3) and (5), we get

$$Bel(A) = Bel(A_0) = m'(A_0) \tag{26}$$

Table 1
Forensic models and feature sets used in the simulations.

Tamper	Forensic model	Feature set	Dimensions
Stego	Provided by paper [37]	X_1	274
Splicing	Provided by paper [20]	X_2	98
Re-compressed	Provided by paper [22]	X_3	144
CG	Provided by paper [36]	X_4	78

According to (10), we get

$$Pl(A) = 1 - \sum_{1 \le i \le L} Bel(A_i) = 1 - \sum_{1 \le i \le L} m'(A_i) \tag{27}$$

According to (8), we have

$$d = Pl(A) - Bel(A) = 1 - \sum_{1 \le i \le L} Bel(A_i) - Bel(A_0) = Bel(A_{L+1})$$
$$= m'(L+1) \tag{28}$$

3.4. DSTM$_3$

DSTM$_3$ is similar to *DSTM*$_2$, except that the feature sets used in *DSTM*$_3$ are all of the combinations of the feature sets used in *DSTM*$_1$. In other words, the feature sets used in *DSTM*$_3$ come from $Com\_Sets(X_1, X_2, \ldots, X_K)$, which is like *DSTM*$_1$. In this way, we enlarge the space of evidence, further reduce the uncertainties and conflicts when using different base forensic models to produce the evidence, and improve the reliability of trustworthiness evaluation.

4. Simulation results and discussions

4.1. Simulation setup

Simulations are conducted on the following feature sets and data sets. We use four forensic models shown in Table 1, and seven classes of training images shown in Table 2. The steganography method used in our simulations is MB2 with an embedding rate of 5%. Three simulations are carried out to evaluate the trustworthiness of digital images using the three models described in Section 3.

In our simulations, we assume that the digital image trustworthiness attribute set is $T = \{stego, splicing, re\text{-}compressed, CG\}$. The evaluation proposition is $A : \{I \textit{ is trustworthy}\}$, where A can be divided into $A = \{A_1, A_2, A_3, A_4\}$. $Com\_Sets(X_1, X_2, X_3, X_4)$ has 15 kinds of feature set combinations, where X_1, X_2, X_3, X_4 are feature sets described in Table 1. Feature set X_1 is composed of 274 merged extended DCT and Markov features as introduced in [37] that are used to detect the multi-class JEPG information hiding, feature set X_2 is composed of 98 Markov based features as introduced in [20] that are used to detect splicing images, feature set X_3 is composed of 144 multiple-counting features formed by histograms of low-frequency DCT coefficients as introduced in [22] that are used to detect double-compression in JPEG images, and feature set X_4 is composed of 78 statistical moments of wavelet sub-band histograms in the DFT domain as introduced in [36] that are used to detect CG images.

The image samples that used in our simulations are described in Table 2. Totally there are 7 classes, each with 1600 images. In detail, the 1600 natural images are downloaded from [48] (Philip Greenspun's online images), where the authors provide lots of diverse natural images with content of indoor, outdoor, people, objects, buildings and so on. The 1600 stego images are generated by an information hiding algorithm (MB2) with the original cover images also downloaded from [48]. The 1600 splicing images are downloaded from [49] (The Columbia Image Splicing Detection Evaluation Dataset). This dataset was created by DVMM at Columbia University for benchmarking blind passive image splicing detection algorithms. The 1600 re-compressed images are generated by a re-compress algorithm with the original images also downloaded from [48]. The 1600 CG images are downloaded from [50] (cgchannel.com), where the authors provide a lot of CG images with many kinds of topics. Two sets of multiple tampering images are also used in our simulations. The 1600 splicing+stego images are generated by hiding information in the splicing images, and the original splicing images are downloaded from [49]. The 1600 re-compressed+steg images are generated by hiding information in the re-compressed images, and the re-compressed images are also generated by a re-compress algorithm with the original images downloaded from [48]. The total number of images is 11,200. The following 3 simulations are carried out on these images, with 1200 × 7 images are training images, and 400 × 7 are test (evaluating) images. Fig. 4 shows some sample images used in our simulations, where (1)–(3) are natural images, (4)–(6) are stego images, (7)–(9) are splicing images, (10)–(12) are re-compressed images, (13)–(15) are CG images, (16)–(18) are both stego and splicing images, and (19)–(21) are both stego and re-compressed images. The following simulation results are obtained from these 21 image samples. We also give the simulation results on all of the 2800 testing images at the end of the next subsection.

Table 2
Training image sets and their download sources.

Image set	Download source	Quantity
Natural	Downloaded from [48]	1600
Stego	Cover images from [48]	1600
Splicing	Downloaded from [49]	1600
Re-compressed	Original images from [48]	1600
CG	Downloaded from [50]	1600
Splicing + stego	Original images from [49]	1600
Re-compressed+stego	Original images from [48]	1600

Table 3
Evaluation results of the *DSTM*$_1$ on sample images in Fig. 4

Image	Real class	$Bel(\neg A)$	$Pl(A)$
(1)	Natural	0.12123847	0.87876153
(2)	Natural	0.11329040	0.88670960
(3)	Natural	0.14103107	0.85896893
(4)	Stego	0.91942269	0.08057731
(5)	Stego	0.91305020	0.08694980
(6)	Stego	0.93336735	0.06663265
(7)	Splicing	0.98069434	0.01930566
(8)	Splicing	0.94620031	0.05379969
(9)	Splicing	0.97803882	0.02196118
(10)	Re-compressed	0.99871381	0.00128619
(11)	Re-compressed	0.99794947	0.00205053
(12)	Re-compressed	0.99876829	0.00123171
(13)	CG	0.98569150	0.01430850
(14)	CG	0.98206101	0.01793899
(15)	CG	0.88233656	0.11766344
(16)	Stego+splicing	0.97792522	0.02207478
(17)	Stego+splicing	0.92159638	0.078403628
(18)	Stego+splicing	0.98147962	0.01852038
(19)	Stego+CG	0.90833909	0.09166091
(20)	Stego+CG	0.91053015	0.08946985
(21)	Stego+CG	0.93629228	0.06370772

4.2. Simulation results based on the three models

The first simulation is carried out on the model *DSTM*$_1$. We use 15 kinds of feature combinations in $Com\_Sets(X_1, X_2, X_3, X_4)$ to train the 11,200 images in Table 2, and evaluate trustworthiness of 2800 images following the steps described in Section 3.2. Table 3 shows the evaluation results of *DSTM*$_1$ on images shown in Fig. 4. The top

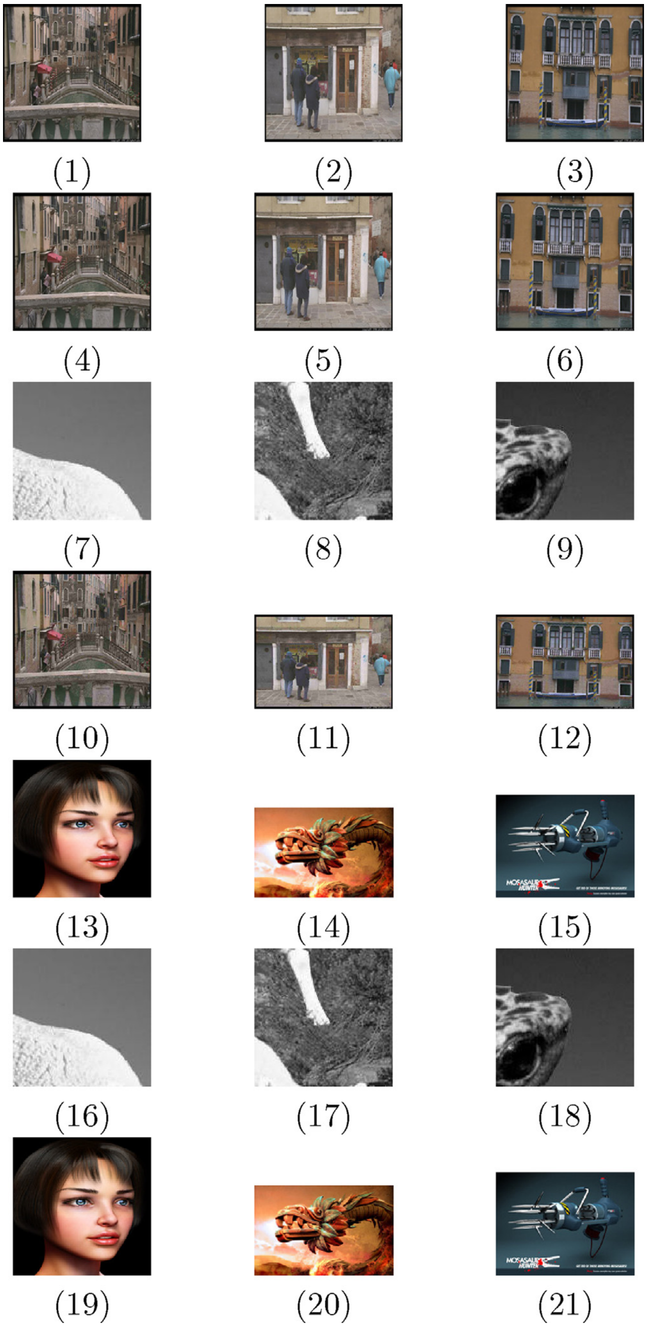


Fig. 4. Some sample of natural and manipulated images used in the simulations.

Table 4
Evaluation results of $DSTM_2$ on sample images in Fig. 4.

Image	Real Class	$Bel(A)$	$Pl(A)$	d
(1)	Natural	0.9983595	0.99843572	0.00007622
(2)	Natural	0.99846499	0.99853857	0.00007358
(3)	Natural	0.99790300	0.99798386	0.00008086
(4)	Stego	0.00073022	0.00079851	0.00006829
(5)	Stego	0.00098687	0.00105813	0.00007126
(6)	Stego	0.00066211	0.00073332	0.00007121
(7)	Splicing	0.00000316	0.00005579	0.00005263
(8)	Splicing	0.00009311	0.00015162	0.00005851
(9)	Splicing	0.00002399	0.00007388	0.00004989
(10)	Re-compressed	0.00000257	0.00004825	0.00004568
(11)	Re-compressed	0.00000521	0.00005120	0.00004599
(12)	Re-compressed	0.00000144	0.00004699	0.00004556
(13)	CG	0.00008738	0.00014304	0.00005566
(14)	CG	0.00013265	0.00018797	0.00005532
(15)	CG	0.00059076	0.07424800	0.00015172
(16)	Stego+splicing	0.00009028	0.00025868	0.00016839
(17)	Stego+splicing	0.00097490	0.11194500	0.00014456
(18)	Stego+splicing	0.00009927	0.00026484	0.00016557
(19)	Stego+CG	0.00125115	0.00132387	0.00007272
(20)	Stego+CG	0.00108686	0.00115562	0.00006876
(21)	Stego+CG	0.00054058	0.00060832	0.00006774

three rows in Table 3 correspond to the three natural images in Fig. 4 without tampering. Rows 4-15 are the evaluation results of images in Fig. 4 with single tampering. Rows 16-21 are the evaluation results of images with multi-tampering (image (16)-(21)). As can be seen, the $Pl(A)$ values of natural images are high, and the $Pl(A)$ values of tampered images are very low. By comparison, the $Pl(A)$ values of multi-tampering are lower than the $Pl(A)$ values of images with single tampering. It shows that our proposed model is effective in evaluating trustworthiness of digital images.

The second simulation is carried out on the model $DSTM_2$. We use 4 feature sets described in Table 1 to train 11200 images described in Table 2, and evaluate trustworthiness of 2800 images following the steps in Section 3.3. Table 4 shows the evaluation results of $DSTM_2$ on the sample images shown in Fig. 4. The last column is the uncertainty defined in Eq. (8). From Table 4, we observe that the $Bel(A)$ and $Pl(A)$ values of the top three natural images (image (1)-(3)) are high. Because the trustworthiness is in the interval of $[Bel(A),Pl(A)]$, we can claim that the top three images in Fig. 4 are trustworthy. We also find the $Bel(A)$ and $Pl(A)$ values of other rows in Table 4 are low, especially for the images with multi-tampering (image (16)-(21)). So, we can claim that the pictures in Fig. 4 with tampering (or multi-tampering) are not trustworthy.

The third simulation is carried out on $DSTM_3$. We use 15 kinds of feature combinations in $Com\_Sets(X_1, X_2, X_3, X_4)$ to train the 11,200 images in Table 2, and evaluate trustworthiness of 2800 images following the steps in Section 3.4. Table 5 shows the evaluation results of model $DSTM_3$ on the sample images in Fig. 4. From Table 5, we find that the $Bel(A)$ and $Pl(A)$ values of the top 3 natural images (image (1)-(3))are high, and the two values in other rows are low. We also find the d values in Table 5 are very low. It means that $DSTM_3$ is very reliable, and can reduce the uncertainty of evidence greatly.

Three models are tested on all of the 2800 images in seven classes. Table 6 shows the mean value of the $Bel(A)$ and $Pl(A)$ of each class of 700 images. The $Bel(A)$ is evaluated by $DSTM_2$ and $DSTM_3$, while the $Pl(A)$ is evaluated by all of the three models. The results

Table 5
Evaluation results of $DSTM_3$ on sample images in Fig. 4.

No.	Real class	$Bel(A)$	$Pl(A)$	d
(1)	Natural	0.99998072	0.99998088	0.00000016
(2)	Natural	0.99998269	0.99998283	0.00000014
(3)	Natural	0.99995846	0.99995868	0.00000022
(4)	Stego	0.00000497	0.00000507	0.00000010
(5)	Stego	0.00000934	0.00000946	0.00000012
(6)	Stego	0.00000385	0.00000396	0.00000011
(7)	Splicing	0.00000001	0.00000004	0.00000003
(8)	Splicing	0.00000029	0.00000036	0.00000006
(9)	Splicing	0.00000004	0.00000007	0.00000003
(10)	Re-compressed	0.00000000	0.00000002	0.00000002
(11)	Re-compressed	0.00000000	0.00000002	0.00000002
(12)	Re-compressed	0.00000000	0.00000002	0.00000002
(13)	CG	0.00000005	0.00000008	0.00000003
(14)	CG	0.00000009	0.00000012	0.00000003
(15)	CG	0.00000062	0.00000086	0.00000024
(16)	Stego+splicing	0.00000023	0.00000059	0.00000037
(17)	Stego+splicing	0.00004304	0.00004447	0.00000143
(18)	Stego+splicing	0.00000060	0.00000124	0.00000064
(19)	Stego+CG	0.00001366	0.00001379	0.00000013
(20)	Stego+CG	0.00000893	0.00090400	0.00000010
(21)	Stego+CG	0.00000267	0.00027600	0.00000009

Table 6
Evaluation results of the three models on 2800 images(7 classes)

Image set	*E*(*Bel*(*A*))		*E*(*Pl*(*A*))		
	$DSTM_2$	$DSTM_3$	$DSTM_1$	$DSTM_2$	$DSTM_3$
Natural	0.99657742	0.90511047	0.99665102	0.99665102	0.90511047
Stego	0.00112020	0.00148629	0.00118838	0.00118838	0.00148629
Splicing	0.00127598	0.00303521	0.00133244	0.00133244	0.00303521
Re-compressed	0.00000517	0.00000025	0.00005205	0.00005205	0.00000025
CG	0.38258628	0.34667702	0.38284205	0.38284205	0.34667702
Stego+splicing	0.00023692	0.00001718	0.00034152	0.00034152	0.00001718
Stego+CG	0.00008678	0.00000522	0.00018727	0.00018727	0.00000522

show that the three models are very stable in evaluating different kinds of images. For the natural images, the *Pl*(*A*) of the three models are high (near to 1); While for the manipulated images, the value are low (near to 0).

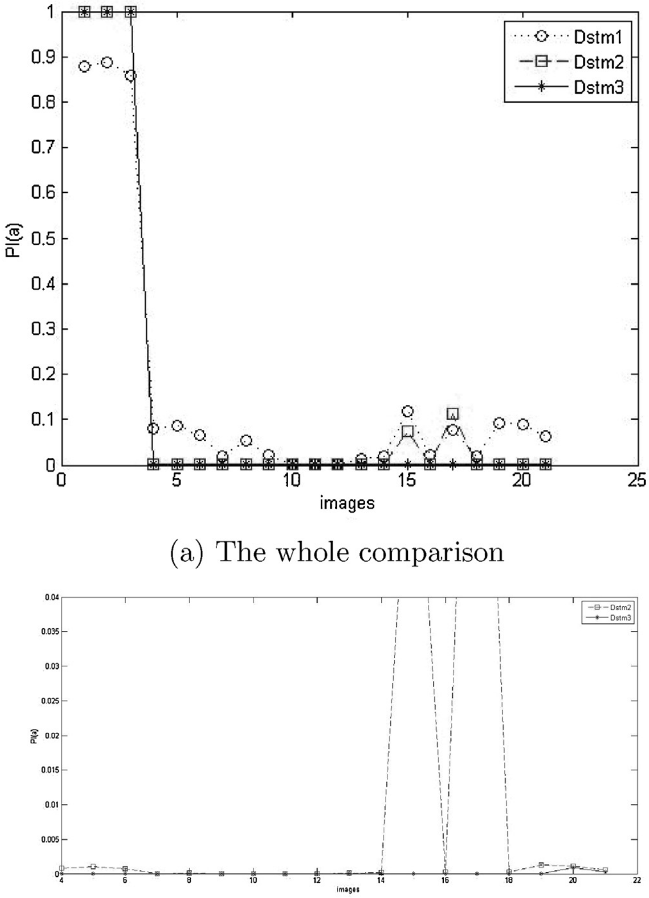
4.3. Comparative studies

Measuring digital image trustworthiness is a new topic introduced in this paper, we compare the performance among our three models. Technically, the digital image forensic method is the most similar to our models, so we also compare digital image trustworthiness evaluation models with digital image forensic models.

Fig. 5 is a comparison of the three digital image trustworthiness evaluation models on images in Tables 3, 4 and 5. From Fig. 5, we can see that, when images are natural (image (1)–(3)), the *Pl*(*A*) values of $DSTM_2$ and $DSTM_3$ are bigger than the *Pl*(*A*) values of $DSTM_1$. When images are tampered (image (4)–(21)), the *Pl*(*A*) values of $DSTM_2$ and $DSTM_3$ are smaller than *Pl*(*A*) values of $DSTM_1$. This is because $DSTM_2$ and $DSTM_3$ use more strict evidence combination rules than $DSTM_1$, and the results of $DSTM_2$ and $DSTM_3$ are more reliable.

Another advantage of $DSTM_2$ and $DSTM_3$ is that the two models can evaluate the uncertainty of the evidence, and provide both lower and upper limits of the trustworthiness. However, $DSTM_1$ can only get the upper limit of the trustworthiness. Fig. 6 shows a comparison of uncertainty evaluation between $DSTM_2$ and $DSTM_3$. We can see that the *d* values from $DSTM_3$ are much lower than that of $DSTM_2$. This is because $DSTM_3$ uses a full feature combination and further reduces the uncertainty.

When digital images have more than one tampering, current digital image forensic methods may fail detecting the real tampering class. Table 7 (α) shows the detection result of applying digital image forensic models to images with multi-tampering (image (16)–(21)) in Fig. 4, from which we can see that the detection results of four forensic models conflict with each other or fail detecting the real tampering class. For example, image (16) is a both stego and splicing image, when using four models (stego, splicing, re-compressed, and CG models) to detect it, as shown in the first four rows in Table 7 (α). The detection results from the four forensic models are stego (with the probability of 0.997102), stego (with the probability of 0.982618), re-compressed (with the probability of 0.899548), and stego (with the probability of 0.906261) respectively. Image (21) is a both CG and stego image, when using four forensic models to detect it, as shown in the last four rows in Table 7 (α). All of the detection results from the four models are re-compressed (with the probability of 0.99997,0.999989,0.337571, and 0.71353 respectively). Using the digital image trustworthiness evaluation methods, we can compute a value that indicates whether the image has been tampered. Moreover, the degree of the tampering (or multi-tampering) can be inferred from the value.



(a) The whole comparison

(b) A zoomed part of the curve in (a)

Fig. 5. The comparison of *Pl*(*A*) among the three models.

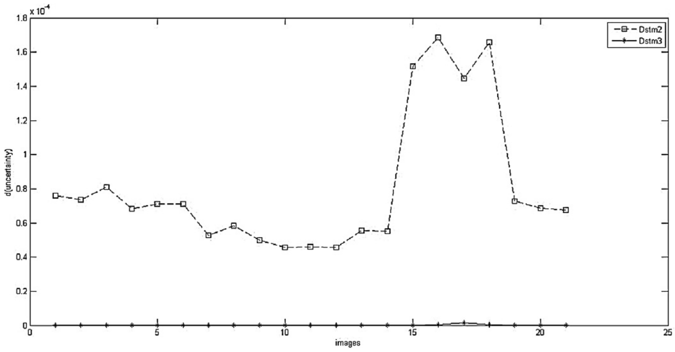


Fig. 6. A comparison of uncertainty between $DSTM_2$ and $DSTM_3$.

Table 7
The results of applying digital image forensic models and digital image trustworthiness evaluation models to images with multi-tampering.

Image	Real class	Forensicmodel	Results of forensic models				
			Natural	Stego	Splicing	Re-compressed	CG
(α) The result of digital image forensic models							
(16)	Stego+splicing	Stego	0.00000255	0.997102	0.000315902	0.000043010	0.00253668
		Splicing	0.0014639	0.982618	0.000891557	0.00142949	0.0137939
		Re-compressed	0.00126674	0.0415238	0.00126633	0.899548	0.0561975
		CG	0.0295152	0.906261	0.00299942	0.020802	0.0404222
(17)	Stego+splicing	stego	0.00161889	0.970634	0.000329978	0.0225959	0.00482099
		Splicing	0.000448011	0.0667226	0.00122231	0.910242	0.0213653
		Re-compressed	0.0297643	0.106906	0.000882658	0.021532	0.840915
		CG	0.110823	0.790452	0.000518767	0.0829419	0.0152647
(18)	Stego+splicing	Stego	9.57E−07	0.998532	0.000131414	0.0000297504	0.00130583
		Splicing	0.000197728	0.0936823	0.000143388	0.892437	0.0135392
		Re-compressed	0.0726689	0.755408	0.000817483	0.123333	0.0477732
		CG	0.00446956	0.976645	4.08E−5	0.00403799	0.0148062
(19)	Stego+CG	Stego	3.47E−8	2.00E−8	1.34E−8	0.99996	3.97E−5
		Splicing	9.56E−8	5.72E−8	3.80E−8	0.999893	0.00010722
		Re-compressed	0.406485	0.00305961	1.83E−6	0.53227	0.058183
		CG	0.30055	8.50E−5	2.56E−6	0.693171	0.00619156
(20)	Stego+CG	Stego	5.74E−8	3.18E−8	2.13E−8	0.999991	8.52E−6
		Splicing	8.61E−8	4.43E−8	2.79E−8	0.999973	2.71E−5
		Re-compressed	0.446852	0.0136247	0.000515057	0.504484	0.0345244
		CG	0.495134	0.000212989	2.51E−6	0.484648	0.0200026
(21)	Stego+CG	Stego	9.88E−8	5.37E−8	3.57E−8	0.99997	2.97E−5
		Splicing	1.11E−8	7.38E−9	4.81E−9	0.999989	1.13E−5
		Re-compressed	0.33516	0.0122055	0.00025822	0.337571	0.314806
		CG	0.276048	0.00125408	1.38E−5	0.71353	0.00915376

Image	Real class	Results of our models($Pl(A)$)		
		$DSTM_1$	$DSTM_2$	$DSTM_3$
(β) The result of digital image trustworthiness evaluation models				
(16)	Stego+splicing	0.00002737	0.00020379	2.488665E−3
(17)	Stego+splicing	0.00026745	0.00059217	1.258904E−0
(18)	Stego+splicing	0.00009013	0.00026854	1.019016E−2
(19)	Stego+CG	0.00070504	0.00076761	1.093986E−0
(20)	Stego+CG	0.00132281	0.00139159	1.447309E−9
(21)	Stego+CG	0.00056908	0.00063759	6.001834E−1

Table 7 (β) shows the $Pl(A)$ values evaluated by digital trustworthiness evaluation models. Apparently, the $Pl(A)$ values of the six images (image (16)–(21)) are (very) low, which means the six images are not trustworthy, or the trustworthiness of the six image are (very) low. No matter what kind of tampering and how many kinds of manipulations an image may have been undergone, our models can evaluate the trustworthiness of the image, while ordinary digital image forensic models may fail predicting it correctly. This difference shows one of the advantages of our digital image trustworthiness evaluation models compared to the traditional digital image forensic models.

Digital image trustworthiness evaluation models can also be used as digital image forensic classification models, although it is not the main objective of this paper. We can classify an image x into a class according to the decision rule as:

$$h(x) = \arg\max_i Bel(A_i) \qquad (29)$$

where $Bel(A_i)$ is calculated by $M'(I)$ in (23) by using the Dempster's rule of combination. And $h(x) = 0$ means that the image x is a natural image, otherwise means that x is a tampered image with tampering i. We compare the forensic accuracy between *DSTM* ($DSTM_2$ and $DSTM_3$) and the four original forensic models, as described in Table 1. The Stego images, splicing images, re-compressed images and CG images used in the simulation are described in Table 2. The comparison results are shown in Fig. 7, where (a), (b), (c) and (d) are detection results of six models (four forensic model of stego, splicing, re-compressed and CG, and our models of $DSTM_2$ and $DSTM_3$) to detect the stego images, the splicing images, the re-compressed images and the CG images respectively. Because each of the base forensic models is designed to capture one side of features, each model can separate a kind of tampering images from natural ones. Meanwhile, it also has the ability to detect other kinds of tampering. For example, the stego model is designed to detect the information hiding in JPEG images. As shown in Fig. 7(a), the stego model can detect stego images with accuracy of 91.5%, but it also can detect splicing images, re-compressed images and CG images with accuracy of 89.25%, 95.75% and 92.75% respectively (as shown in Fig. 7(b)–(d)). Other forensic models may be good at detecting some certain kind(s) of tampering, but fail to detect other kinds of tampering. For example, from Fig. 7(b) and (c), we can see that the re-compressed model can detect splicing images and re-compressed images with accuracy of 88.25% and 89.25% respectively. When detecting the tampering of stego and CG (as shown in Fig. 7(a) and (d)), it only can achieve the poor accuracy of 38.25% and 47.00%. Our method (DSTM2 and DSTM3) can combine the ability of four base forensic models together and at the same time reduce the evidence conflict by using information fusion and the D-S theory. As a result, our method can detect all of the four kinds of tampering with very high detection accuracy. For example, as shown in Fig. 7(a)–(d), DSTM2 can detect stego images, splicing images, re-compressed images and CG images with very high accuracy of 99.5%, 98.75%, 100% and 100% respectively.

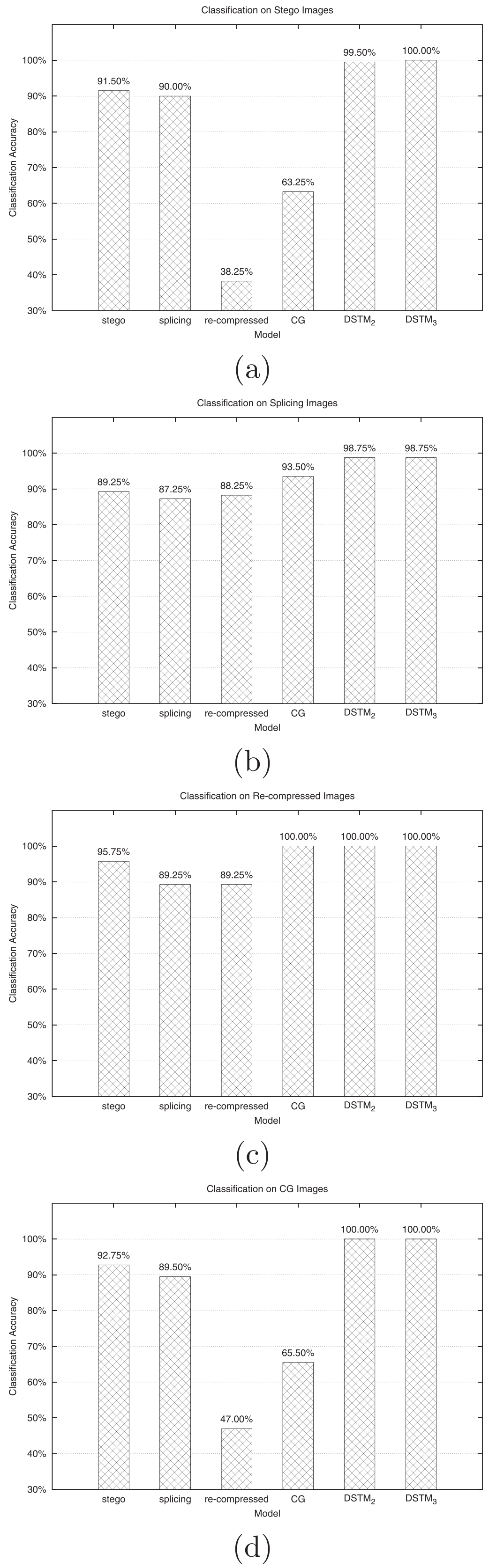


Fig. 7. A comparison of image forensic accuracy between *DSTM* and other image forensic models.

5. Conclusions and future work

Digital images are facing a crisis of trustworthiness. This paper proposed a method to evaluate digital image trustworthiness. Our method is different from existing digital image forensics as our method does not need the help of digital watermarking or image hash. We proposed three models based on the D-S evidence theory and data fusion. Among them, $DSTM_1$ directly uses the D-S theory to evaluate the *Pl* value of trustworthiness, and uses the feature combinations to resolve the problem of evidence combinations. $DSTM_2$ uses the least square method to compute the evidence matrix before directly using the Dempster's rule of combination. The uncertainty degree can also be evaluated in this model. $DSTM_3$ is a combination of $DSTM_1$ and $DSTM_2$. Compared to $DSTM_2$, $DSTM_3$ improves the evaluation reliability, and also greatly reduces the uncertainty by using the feature combination method. Simulation results have shown that the three models are effective in evaluating trustworthiness of digital images. The main contributions of this work are summarized as follows:

1 We proposed a new approach that can measure the trustworthiness of a digital image. The trustworthiness of a digital image is a reference value to tell the degree of trustworthiness of the image based on the trustworthiness evaluation indexes a user selected according to her application requirements. Typical application scenarios include online digital image retrieval, image fusion, online social network digital image sharing, etc. To the best of our knowledge, our work is the first research on how to measure the trustworthiness of digital images in a quantitative way.
2 We proposed three digital image trustworthiness evaluation models based on information fusion for measuring the trustworthiness of a digital image. Experimental results demonstrated that the three models are reliable and effective in evaluating the trustworthiness of natural images.
3 We applied the D-S theory at the feature fusion level, the decision fusion level, and the combination level of feature fusion and decision fusion, respectively, to calculate the trustworthiness value of a digital image, and at the same time compute the uncertainties introduced by base forensic models. In addition, an improved least square method was proposed (in the second and third models) to reduce conflicts among forensic evidence provided by different forensic models.
4 The proposed digital image trustworthiness evaluation models can be adapted as digital image forensic classification models (although it is not the main objective of this paper), and the experimental results demonstrated that when being adapted as digital image forensic models, our models can improve the forensic accuracy significantly compared to the original base forensic models.

Meanwhile, our approach still has some weaknesses. First, our approach depends on a base forensic model to capture each kind of detailed tampering, and then uses a unified framework to combine the base forensic models together. Our approach assumes that all of the forensic models can use machine learning methods to predict a tampering by previously trained classifiers, so it is hard to incorporate some forensic models based on some special or distinct statistics, such as individual statistics of sensor dust [34] or inconsistencies in lighting [51]. In the future, we plan to develop new statistical methods that can "compute" the trustworthiness of a digital image on the tampering statistics directly (using other kinds of statistical models). Second, although our approach is different from traditional digital image forensic methods, it cannot replace the digital image forensic methods themselves. Instead, we still need to develop each kind of digital image forensic model to accurately capture the feature or evidence of a certain kind of tampering. When a digital image undergoes many kinds of tampering, although our approach can measure its trustworthiness while current forensic models may fail to accurately detect multiple tampering, we still need to develop new forensic methods to detect multiple tampering for real-world digital image applications, or even to detect the processing ordering of multiple tampering. Third, in this work, when calculating the trustworthiness of a digital image, we only consider the (tampering) feature of the digital image itself, while in real-world applications, we can further consider other elements, such as contextual information, human-system interaction information, and so on. For example, when evaluating the trustworthiness of a digital image which is shared in an Online Social Network (OSN), we can consider the relationship of the image's owner and the person whom the image is shared to, as well as the comments which the shared image received from other OSN users. A more challenging issue is whether and how we can take into account semantic meanings of a digital image (e.g., people or objects appeared in the image, time and location it is taken, and relationships between the people or objects appeared in the image and the individual who posts the image in the OSN or the individual who claims as the image's owner) when evaluating its trustworthiness. We plan to investigate this issue in our future work.

Our trustworthiness evaluation approach proposed in this paper is based on passive blind image forensic methods. Active image forensic methods, such as perceptual image hash [6,7,52], can also be used in digital image trustworthiness assurance. How to design methods that can evaluate the trustworthiness of a digital image based on perceptual image hash is another direction of our future research work.

While the trustworthiness evaluation is to tell to what extent a digital image is trustworthy (to be used or shared), a somewhat opposite direction is whether we can compute the degree of the images privacy. There are many scenarios where the privacy degree of digital images can be used. For example, an OSN user can use the privacy degree as a reference for privacy decision before she shares the image, and an OSN access control system can use it as useful reference information too. In our future work, we also plan to investigate digital image privacy issues in OSNs.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful and constructive comments and suggestions that have helped improve the quality of this paper. This research has been supported by the National 973 Program of China (Grant 2013CB329604), the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China (Grant IRT13059), and the National Natural Science Foundation of China (Grant Nos. 61272540, 61229301 and 61375047).

References

[1] S.P. Marsh, Formalising trust as a computational concept (Ph.D. thesis), Department of Computing Science and Mathematics, University of Stirling, 1994.
[2] K. Dirks, The effects of interpersonal trust on work group performance, J. Appl. Psychol. 84 (3) (1999) 445–455.
[3] Y.Y. Yamamoto, A morality based on trust: some reflections on Japanese morality, Philos. East West 40 (4) (1990) 451–469.
[4] T. Grandison, M. Sloman, A survey of trust in internet applications, IEEE Commun. Surv. Tutor. 3 (4) (2000) 2–16.
[5] I.J. Cox, M.L. Miller, J.A. Bloom, Digital Watermarking, Morgan Kaufmann, 2001.
[6] A. Swaminathan, Y. Mao, M. Wu, Robust and secure image hashing, IEEE Trans. Inf. Forensics Secur. 1 (2) (2006) 215–230.
[7] L. Wang, X. Jiang, S. Lian, D. Hu, D. Ye, Image authentication based on perceptual hash using gabor filters, Soft Comput. 15 (3) (2011) 493–504.
[8] H. Sencar, N. Memon, Overview of state-of-the-art in digital image forensics, Algorithms Archit. Inf. Syst. Secur. 3 (2008) 325–348.
[9] D. Hu, L. Wang, Y. Zhou, Y. Zhou, X. Jiang, L. Ma, D–S evidence theory based digital image trustworthiness evaluation model, in: International Conference on Multimedia Information Networking and Security, (MINES '2009), Vol. 1, Hubei, China, 2009, pp. 85–89.
[10] A. Dempster, Upper and lower probabilities induced by multivalued mapping, Ann. Math. Stat. 38 (2) (1967) 325–339.
[11] S. Glenn, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[12] L. Thanuka, P. Kamal, K. Miroslav, T.J. Dushyantha, A morality based on trust: some reflections on Japanese morality, IEEE Trans. Knowl. Data Eng. 23 (2) (2011) 175–189.
[13] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recogn. Lett. 24 (13) (2003) 2115–2125.
[14] A. Gunatilaka, B. Baertlein, Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 577–589.
[15] M. Johnson, H. Farid, Exposing digital forgeries by detecting inconsistencies in lighting, in: 7th workshop on Multimedia and security, New York, NY, USA, 2005, pp. 1–10.
[16] M. Johnson, H. Farid, Exposing digital forgeries through specular highlights on the eye, in: Information Hiding, Vol. 4567 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg/Saint Malo, France, 2007, pp. 311–325.
[17] J. Lukáš, J. Fridrich, M. Goljan, Detecting digital image forgeries using sensor pattern noise, in: SPIE Electronic Imaging, San Jose, CA, USA, 2006, p. 60720Y.
[18] M. Johnson, H. Farid, Exposing digital forgeries through chromatic aberration, in: 8th workshop on Multimedia and security, New York, NY, USA, 2006, pp. 48–55.
[19] T. Ng, S. Chang, Q. Sun, Blind detection of photomontage using higher order statistics, in: International Symposium on Circuits and Systems, (ISCAS'04), Vol. 5, Sheraton Vancouver Wall Centre Hotel, Vancouver, Canada, 2004, pp. 688–691.
[20] Y. Shi, C. Chen, W. Chen, A natural image model approach to splicing detection, in: 9th Workshop on Multimedia & Security, Dallas, TX, USA, 2007, pp. 51–62.
[21] J. Lukáš, J. Fridrich, Estimation of primary quantization matrix in double compressed jpeg images, in: Digital Forensic Research Workshop, Cleveland, USA, 2003, pp. 5–8.
[22] T. Pevny, J. Fridrich, Detection of double-compression in jpeg images for applications in steganography, IEEE Trans. Inf. Forensics Secur. 3 (2) (2008) 247–258.
[23] T. Pevn'y, J. Fridrich, Estimation of primary quantization matrix for steganalysis of double-compressed jpeg images, in: SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents, San Diego, CA, United States, 2008, pp. 11–24.
[24] R. Böhme, M. Kirchner, Counter-forensics: attacking image forensics, in: Digital Image Forensics, Springer, New York, 2013, pp. 327–366.
[25] G. Cao, Y. Zhao, R. Ni, X. Li, Contrast enhancement-based forensics in digital images, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 515–525.
[26] A. Popescu, H. Farid, Exposing digital forgeries by detecting traces of resampling, IEEE Trans. Signal Process. 53 (2) (2005) 758–767.
[27] M. Kirchner, T. Gloe, On resampling detection in re-compressed images, in: First IEEE International Workshop on Information Forensics and Security, WIFS 2009, London, UK, 2009, pp. 21–25.
[28] M. Stamm, K. Liu, Blind forensics of contrast enhancement in digital images, in: 15th IEEE International Conference on Image Processing, ICIP 2008, San Diego, CA, 2008, pp. 3112–3115.
[29] C. Chen, J. Ni, J. Huang, Blind detection of median filtering in digital images: a difference domain based approach, IEEE Trans. Image Process. 22 (12) (2013) 4699–4710, http://dx.doi.org/10.1109/TIP.2013.2277814.
[30] Y. Zhang, S. Li, S. Wang, Y.Q. Shi, Revealing the traces of median filtering using high-order local ternary patterns, IEEE Signal Process. Lett. 21 (3) (2014) 275–279.
[31] K. Mehdi, H. Sencar, N. Memon, Blind source camera identification, in: International Conference on Image Processing, Vol. 1, Singapore, 2004, pp. 709–712.
[32] S. Bayram, H. Sencar, N. Memon, I. Avcibas, Source camera identification based on CFA interpolation, in: IEEE Int. Conf. Image Processing, Vol. 3, Magazzini de Cotone Modulo 9, Genova, Italy, 2005, pp. 69–72.
[33] Z. Geradts, J. Bijhold, M. Kieft, K. Kurosawa, K. Kuroki, N. Saitoh, Methods for identification of images acquired with digital cameras, in: Proc. SPIE 4232, Enabling Technologies for Law Enforcement and Security, 2001, pp. 505–512.
[34] A.E. Dirik, H.T. Sencar, N. Memon, Source camera identification based on sensor dust characteristics, in: IEEE Workshop on Signal Processing Applications for Public Security and Forensics, Washington, DC, USA, 2007, pp. 1–6.
[35] S. Lyu, H. Farid, How realistic is photorealistic? IEEE Trans. Signal Process. 53 (2) (2005) 845–850.
[36] X. Cui, X. Tong, G. Xuan, C. Huang, Discrimination between photo images and computer graphics based on statistical moments in the frequency domain of histogram, in: Seventh China Information Hiding Workshop, Nanjing, 2007, pp. 276–283.
[37] T. Pevn'y, J. Fridrich, Merging Markov and DCT features for multi-class jpeg steganalysis, in: SPIE Electronic Imaging Photonics West, San Diego, CA, United States, 2007, pp. 3–4.

[38] J. Fridrich, M. Goljan, D. Soukal, T. Holotyak, Forensic steganalysis: determining the stego key in spatial domain steganography, in: SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VII, San Jose, CA, USA, 2005, pp. 631–642.

[39] J. Fridrich, M. Goljan, D. Soukal, T. Holotyak, Searching for the stego-key, in: SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VI, Vol. 5306, San Diego, CA, USA, 2004, pp. 70–82.

[40] M. Barni, A. Costanzo, Dealing with uncertainty in image forensics: a fuzzy approach, in: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP '2012), Kyoto, Japan, 2012, pp. 1753–1756.

[41] D. Zhang, Inconsistencies in information security and digital forensics, in: IEEE International Conference on Information Reuse and Integration, (IRI '2010), Las Vegas, USA, 2010, pp. 141–146.

[42] H. Farid, Seeing is not believing, IEEE Spectr. 46 (8) (2009) 44–51.

[43] B. Mahdian, S. Saic, Using noise inconsistencies for blind image forensics, Image Vis. Comput. 27 (2009) 1497–1503.

[44] Z.-Q. Zhao, H. Glotin, Z. Xie, J. Gao, X. Wu, Cooperative sparse representation in two opposite directions for semi-supervised image annotation, IEEE Trans. Image Process. 21 (9) (2012) 4218–4231.

[45] Y. Freund, R. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, Vol. 904 of Lecture Notes in Computer Science, 1995, pp. 23–37.

[46] X. Yin, C. Liu, Z. Han, Feature combination using boosting, Pattern Recogn. Lett. 26 (14) (2005) 2195–2205.

[47] T. kuo Huang, R.C. Weng, C. jen Lin, G. Ridgeway, Generalized Bradley–Terry models and multi-class probability estimates, J. Mach. Learn. Res. 7 (7) (2006) 85–115.

[48] P. Greenspun, Philip greenspun's online images', http://philip.greenspun.com/images.

[49] C.D.R. Lab, Dvmm – demos and downloads, http://www.ee.columbia.edu/ln/dvmm/newdownloads.htm.

[50] Cg channel, http://www.cgchannel.com/category/cgelite.

[51] E. Kee, J.F. O'brien, H. Farid, Exposing photo manipulation from shading and shadows, ACM Trans. Graph. 33 (5) (2014) 165:1–165:21.

[52] Y. Zhao, S. Wang, X. Zhang, H. Yao, Robust hashing for image authentication using zernike moments and local features, IEEE Trans. Inf. Forensics Secur. 8 (1) (2013) 55–63.