Possible counter-attacks against random geometric distortions

Jean-Luc Dugelay[a] and Fabien A. P. Petitcolas[b]

[a] Institut Eurécom, France, Jean-Luc.Dugelay@eurecom.fr
[b] Microsoft Research, fabienpe@microsoft.com

ABSTRACT

After a brief reminder on the real difficulties that digital watermarking software still has to tackle – especially some random geometric attacks such as StirMark[1] – we present an early overview of on-going solutions to make the survival of the watermark possible.

1. TACKLE THE REAL PROBLEMS

The recent literature on digital watermarking has introduced a new paradigm where signal processing, computer security, cryptography, law and business converge to protect the rights of photographers, digital artists, singers, composers, in short, copyright holders. The view of a senior executive at a California-based record label is significant of the expectations surrounding digital watermarking: '*Sooner or later, any encryption system can be broken. We need watermarking technologies to tell us who did it*'[2].

In this Wonderland, Alice*, the copyright holder, sells an object to Mallory who redistributes it further. Fortunately Alice has embedded a strong digital watermark, a digital fingerprint and a fragile watermark in the object and she uses an automated Web-based copyright audit system. This system uses the strong watermark to detect illegal copies and reports any infringement to Alice, who then extracts the fingerprint. Alice now has all the material to sue Mallory: She can prove that she owns the copyright (using the strong watermark); she can prove that the image has been modified (using the fragile watermark); and she can prove that the culprit is Mallory (using the fingerprint).

Many attacks, but in particular random geometric distortions[1,3], have shown that the state-of-the-art is still far from achieving what has been promised by the industry: Lack of standardisation, interoperability issues, lack of a set of precise and realistic requirements for watermarking systems and lack of results on important issues, still severely hinder the development of copy protection mechanisms. Strong watermarking and fingerprinting may exist but until this is proved, these attacks show the contrary.

Meanwhile we should also consider the fact that the attacker is a person who may have access to future technology – for example, a pirate seeking to remove the watermark or fingerprint embedded in a 1997 music recording using the technology available in 2047. This is a serious concern with copyright, which may subsist for a long time (typically 70 years after the author's death for text and 50 years for audio). Even where we are concerned only with the immediate future, Gurnsey notices from industry experience, that it is a 'wrong idea that high technology serves as a barrier to piracy or copyright theft; one should never underestimate the technical capability of copyright thieves'[4]. Such experience is emphasised by the recent success of criminals in cloning the smartcards used to control access to satellite TV systems[5] and the recurrent failures of 'magic technologies'[6] against piracy.

Figure 1 summarises the general watermarking setting and its main challenges. An owner would like to protect their image rights. The original image is denoted *I*. To do so, they add a mark *M* to the image (hopefully) without introducing any visual

* In the security literature parties are usually referred to as Alice (A) and Bob (B). The attacker, depending on their role, may be called Mallory.

degradation. The protected image is denoted $\tilde{I}$. When needed, they would like to prove their ownership, by retrieving the watermark despite possible modifications of the image. The corrupted version of the image $\tilde{I}$ is denoted $\tilde{I}'$. The operations of insertion, extraction and corruption are respectively denoted by the '+', '*' and '–' operators.

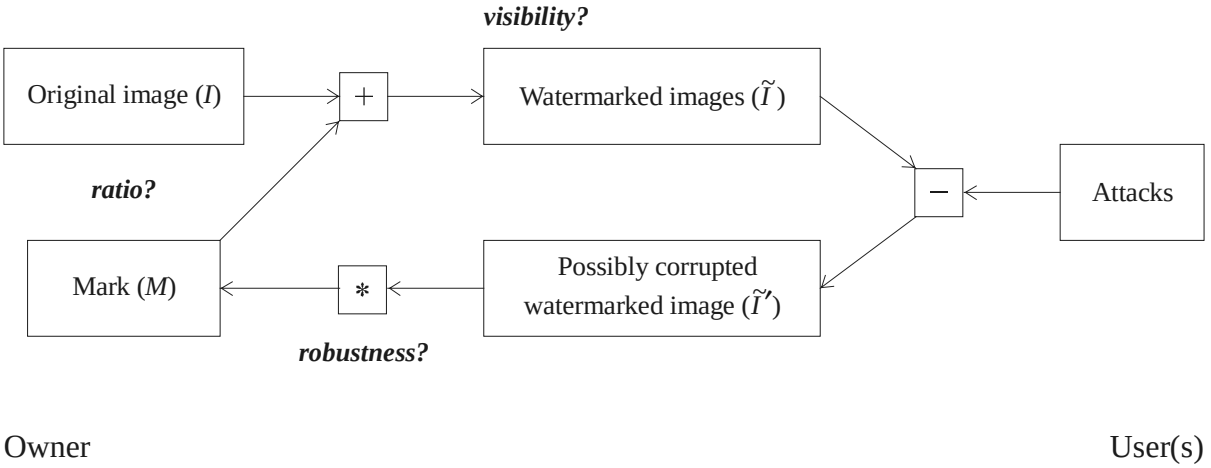


Figure 1 – Basic scheme of invisible robust digital watermarking.

Three aspects have to be considered:

- The ratio between the information contained in the watermark and the information conveyed by the image;
- The image degradation due to watermarking;
- And the 'robustness' to 'non-destructive' attacks.

In addition, several modes of extraction exist (blind, semi-blind or non-blind), depending on the necessity to use original information (i.e. the host signal and/or the watermark) but blind is the most challenging and has more applications.

2. DIFFICULT ATTACKS

The recent research on digital watermarking has emphasised three basic rules that can help to have better robustness.

Spreading the mark. Maybe the analogue of the 'diffusion' principle used in cryptography. This ensures that any part of the cover-signal contains the hidden data and avoids removal by simple band filtering or cropping.

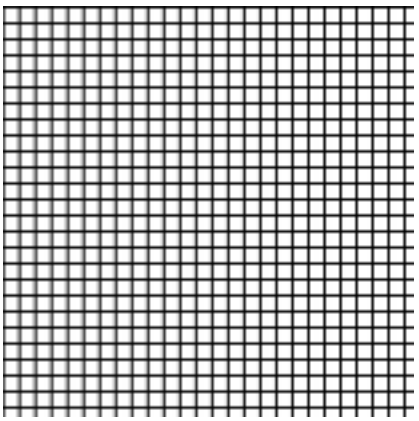
Random selection of features in a transform space. Perhaps the analogue of 'confusion' in cryptography. This makes sure that the attacker cannot find out where the information actually is and the use of the key ensures that only authorised people can access the mark. The hiding process does not need to take place on the samples themselves if transform spaces give advantages. Actually transform spaces have been used very often to give better results against compression: A mark hidden in a particular transform space is expected to survive better compression based on the same transform space (e.g., D.C.T.-JPEG).

Masking. Take explicit advantage of the properties of the human visual or auditory systems. Early systems are basically heuristic in the sense that experiments have shown that they do not introduce annoying distortions in the image. Use of masking ensures that the strength of the mark is as high as possible but imperceptible and can typically be summarised like this: The mark is modulated by the output of a perceptual model and then added in a transform space.

The first property stems from many variants of spread-spectrum modulation used for digital watermarking. Unfortunately most simple spread-spectrum based techniques – hence most current proposals – are still subject to some kind of jitter attack[3]. Indeed, although spread-spectrum signals are very robust to amplitude distortion and to noise addition, they do not survive timing errors: synchronisation of the chip signal is very important and simple systems fail to recover this synchronisation properly. So one way to attack such systems is to break up the synchronisation needed to locate the samples in which the mark is hidden.

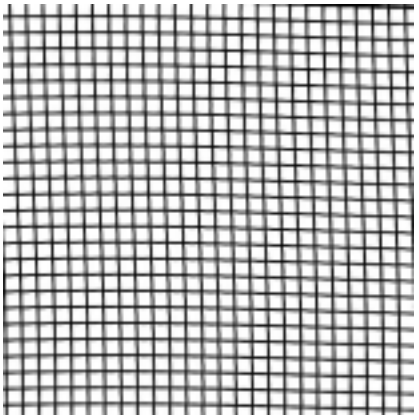
Assuming a binary shift-register sequence s of length N is used as spreading sequence, the autocorrelation with lag l of this sequence is $R_s(l) = A_l - D_l$ where A_l denotes the number of places in which the vectors s and s-shifted-by-l-places agree, and $D_l = N - A_l$ is the number of places in which they disagree. Suppose that one value is removed at index j in s and the sequence is padded with 0, producing a new sequence n' of length N. It is clear that the smaller j the more the 0-lag cross-correlation between s and s' will be decreased. However the 1-lag correlation will be very good. So $j = N/2$ gives best results in average as in this case the correlation peak cannot be greater than $N/2$. More generally, using regular interval will give better results in average than random one. Unfortunately it is not very easy to characterise this phenomenon but the effect is: extra correlation peaks and weaker 0-lag correlation. To be efficient the attack should be applied in the domain in which the watermarking algorithm works. Unfortunately almost no work has been done to study partial correlation of random sequences and we are aware of only one paper dealing with this problem[7]. The best you can do is to resynchronise every so often.

(a)

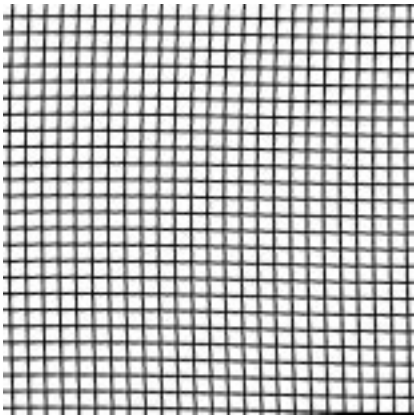


(b)

(c)



(d)



(e)

Figure 2 – Illustration of StirMark's random geometric distortions. (a) and (b) are two 512×512 original images. StirMark 3.1's distortions have been applied to them and the results are (c), (d) and (e). The same distortions have been applied to both (c) and (d). To illustrate the fact that the process is random (d) and (e) have been generated from (b) using the same StirMark command line.

Using this idea a first trivial attack you can do is to simply remove lines or columns of pixels in an image. Elaborating on this idea, it appears that with small random geometric distortions, it is possible to defeat many proposed schemes. Examples of such distortions are the one applied by StirMark[1]. StirMark 1.0 applied a bilinear transformation to the picture by moving its corners by a small random amount. Since version 1.3 a highly non-linear displacement is also added to each pixel: A kind of embossment with imperceptible random wobbling. It is worth noting that these geometric distortions are also combined with a fast resampling[8] and a mild JPEG compression. Figure 2 shows that these distortions are almost invisible when applied to photographs and also shows what these distortions look like. This can also be applied to video, provided that the random parameters are saved and used for all the frames in the same sequence. The natural extension of these ideas is to add a wobbling at each frequency whose amplitude is inversely proportional to the frequency, leading to transformations with, theoretically, an infinite degree of freedom. This is what is being implemented in StirMark 4.0.

3. SOLUTIONS

This difficult problem of geometrical attacks led us to suggest that detection, rather than embedding, is the core problem of digital watermarking. But it seems that the set of issues around mapping back the home space of the mark after an attacker has tampered with it, and finding a metric that encompasses all valuable objects in the vicinity is still really underestimated.

Some solutions propose to use the original image (or some kind of information pre-extracted from it) in order to estimate the distortions, say f, and then try do detect the mark in $f^{-1}(\tilde{I}')$ expecting that the distortions will be compensated. Although this helps to counter-attack geometrical distortions, this has only limited application since it is non-blind. Another idea, which will also be discussed later, is to use a reference in the cover-image. This reference can be inherent to the image but can also be added together with the mark. At last, if all these tricks fail, the brute force search remains.

3.1. NON-BLIND

The solution proposed by Davoine et al.[9] has been inspired by previous work done on video using triangular patches for motion compensation[10]. The basic idea is to split the original (or uncorrupted watermarked) picture into a set of triangular patches. This mesh will then serve as reference mesh and be kept into memory for a pre-processing step of watermark retrieval.

A similar splitting is then performed on the corrupted, watermarked image. Using a compensation procedure, which allows affine transformations, an approximate inversion attack transformation is computed. When this is done, the retrieval of the watermark can be done correctly in many cases.

As emphasised by the authors, this kind of compensation based on active meshes is only efficient in case of minor deformations (such as the ones due to the motion of some objects in a video between two consecutive frames). In case of major local deformations or local deformations combined with some global attacks on the image, the compensation can provide to some false matching between meshes and then does no longer improve the extraction of the watermark by the retriever.

In order to reduce the size of data kept into memory, the authors propose to use a binary map of edges instead of the whole picture. Moreover, as a perspective, they envisage deriving the pre-processing step of geometric compensation via a limited set of feature pixels. If the retriever were able to identify alone these features from the corrupted watermarked image, a blind extraction mode would then be possible.

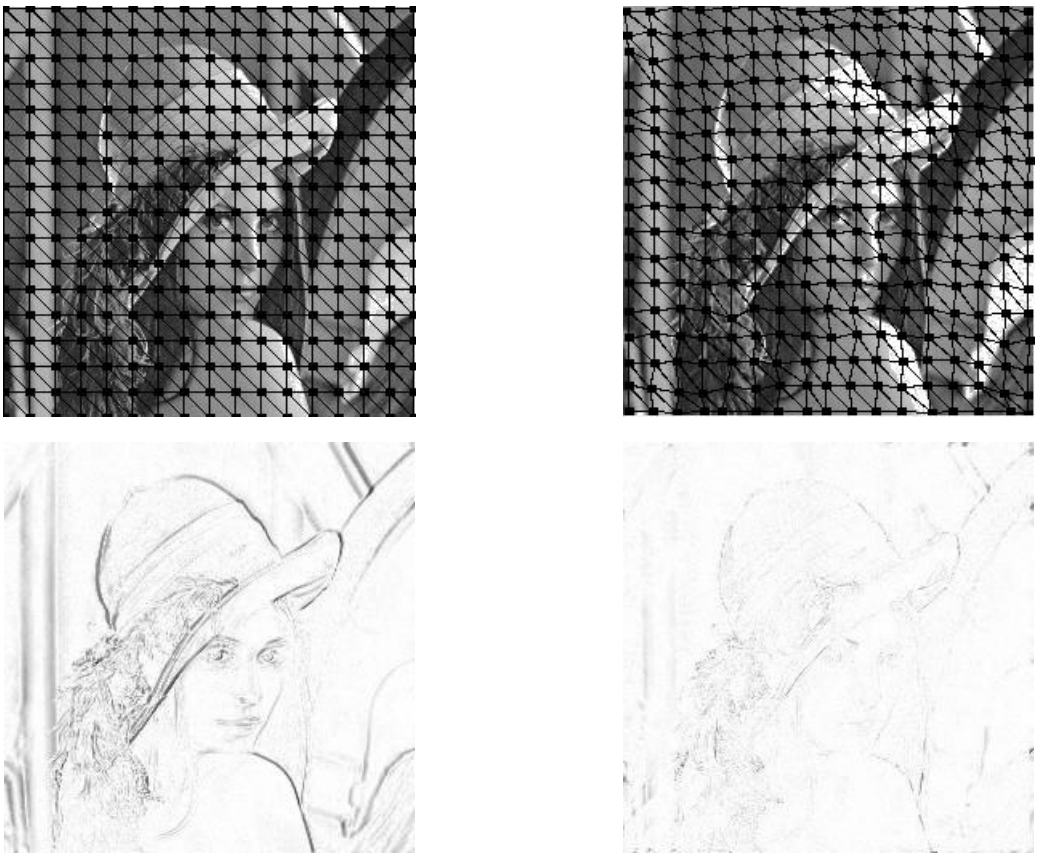


Figure 3 – Original image/mesh (a). Corrupted, watermarked image/mesh (b). Difference between original and corrupted, watermarked image, before compensation (c). Difference between original and corrupted, watermarked image, after compensation (d). Courtesy of F. Davoine.

Actually this is what Johnson et al.[11] propose for a method to invert affine transformations (defined by 6 unknown parameters) of images using the original image. Here the 'feature points' are groups of neighbouring pixels (3×3 to 11×11) with very low correlation with any other group of pixels nearby (60×60) or neighbouring pixels with high 'cornerness' measure. The feature points are extracted in the modified image and in the original. The best affine transformation – in the least square sense – can then be estimated by using the correspondences between image points in the original and transformed images.

3.2. BLIND

In order to preserve the possibility to retrieve the watermark without requiring the use of any original information, some alternatives have been proposed by Kutter[12]. A first concept of symbol reference has been introduced. The basic idea is to pre-set parts of the watermark to known values and to use them for spatial resynchronisation. But, this approach has the disadvantage to decrease the hiding ratio since only a limited part of the watermark (useful bits) includes some information; other bits are dedicated to resynchronisation. Moreover, even if it is clear that any affine geometric transformations (so only 6 degrees of freedom) could possibly be compensated by testing all possible inverse transformations and selecting the one giving best results, this approach is computationally very expensive when considering full images.

So, instead of using feature points which have to be compared before and after attack in order to compute the inverse transformation (hence requiring a non blind mode of extraction), Kutter introduced the idea of self-reference systems that embed

the watermark several times at shifted locations. The watermark becomes a reference itself, making the synchronisation possible without using original information, simply using the relative position of the marks.

The methods overviewed until now consider the image as a whole and severely reduce the space of distortions they look at (typically 6 dimensions), providing robustness to the so called 'generalised geometric transformations' (see Figure 4), that is affine transformations (i.e., simple combination of scaling, rotation, shearing and shifting). This is a first promising step towards robustness against more general distortions, which, as said earlier, may introduce many more degrees of freedom – hopefully an infinity using fractal distortions.

Figure 4 – A typical 'generalised geometric transformation'.

So, to better deal with the complexity of the possible attacks one may try to inverse the distortions only locally and hence use a bloc based approach. Indeed, although the global modifications can be very high and highly non linear, the local one (e.g., on 10×10 pixel blocks) are very minor because the attacks must preserve, to some extent, the visual quality of the image. So locally the distortions can better be approximated by an affine linear transformation.

Hartung et al.[13] use this idea to improve their spread-spectrum based watermark retrieval technique. The typical correlation computation used in the detection process is applied to blocks, where it is maximised with respect to the original pseudo-noise signal used for embedding. More precisely, if A is the set of affine transformations which is searched, then, for each image block B_i, they try to maximise something like:

$$corr = \max_{f \in A} \left\langle f\left(B_i\right), s_j \right\rangle$$

where s_j is the spreading sequence.

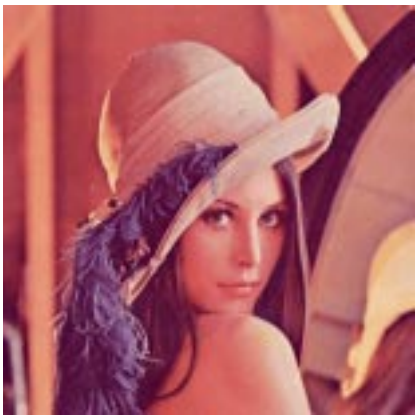
The advantage of the block-based approach is that, although the searched space has limited dimension, the over whole searched space is much larger that the previously discussed global approaches.

4. CONCLUDING REMARKS

As seen, robustness against random geometric attacks is not a trivial task. Nevertheless, new techniques appear. They are generally applied as a pre-processing step of the retrieval procedure. Some work with the original image or at least with the knowledge of some image features, but other propose some solutions to work via a full blind mode of extraction and new

encouraging results (see Figure 5) have recently been published[14] and show that skilful combination of the paradigms described in section 3 together with efficient hiding spaces (e.g., self-similarities[15]) may solve the current random geometric distortions problem until the next generation of attacks becomes available.

It is hard to tell to which degree of distortions a marking method ought to survive. So robustness evaluation benchmark[16] could use ideas of the evaluation of security systems[17]. Different levels of robustness corresponding to different sets of attacks and applications should be defined. The larger the set, the better the robustness. It is time now to finally agree on a minimal definition of 'robustness'.



Watermarked image (distortion 38.08 dB)
Secret key: '1234'. Watermark: 'IEEE'

Corrupted, watermarked image. Size 538×538
Attacks: StirMark 3.1 + horizontal flip + 3° rotation

Figure 5 – In this example, the size of the hidden message is 32 bits (4 characters). The marking process is performed with very little visual alteration, and is secured by a 4-digit secret key. Using this key, the mark can be recovered in a robust and completely blind manner from the attacked watermarked document. See Figure 2(a) for original image.

ACKNOWLEDGEMENTS

Many thanks to Thierry Pun (University of Geneva), Frank Davoine (Université de Technologie de Compiègne), Cormac Herley (Microsoft Research), Stéphane Roche (Institut Eurécom), and Christian Rey (Institut Eurécom) for useful discussions.

REFERENCES

1 <http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/>.

2 J. Yoshida, 'Digital watermarking showdown between ARIS and Blue Spike'. EETimes on-line, 3 Jun. 1999, quotation of a senior executive at a California-based record label, <http://www.eetimes.com/story/OEG19990603S0034>.

3 F. A. P. Petitcolas, R. J. Anderson and M. G. Kuhn. 'Attacks on copyright marking systems'. In D. Aucsmith, ed., *Information hiding: second international workshop*, vol. 1525 of Lecture notes in computer science, Portland, Oregon, U.S.A., pp. 218–238, Apr. 1998, Springer Verlag, Berlin, Germany. ISBN 3-540-65386-4.

4 J. Gurnsey, *Copyright theft.* Aldershot, England: Aslib Gower, 1995, ISBN 0-566-07631-4.

5 R. J. Anderson and M. G. Kuhn, 'Tamper resistance – a cautionary note'. In *second USENIX workshop on electronic commerce*, pp. 1–11, Oakland, California, U.S.A., 18–21 Nov. 1996, ISBN 1-880446-83-9.

[6] B. Fox, ‘The pirate’s tale’. *The New Scientist*, no. 2217, pp. 40–43, 18 Dec. 1999, <http://www.newscientist.com/ns/19991218/thepirates.html>.
[7] D. V. Sarwate, M. B. Pursley and T. Ü. Basar, ‘Partial correlation effects in direct-sequence spread-spectrum multiple-access communication systems’. *IEEE transactions on communications*, vol. COM-32, no. 5, pp. 567–573, May 1984.
[8] N. A. Dodgson, ‘Quadratic interpolation for image resampling’. *IEEE transactions on image processing*, vol. 6, no. 9, pp. 1322–1326, Sep. 1997, ISSN 1057-7149.
[9] F. Davoine, P. Bas, P.-A. Hébert, and J.-M. Chassery, ‘Watermarking et résistance aux déformations géométriques’. In J.-L. Dugelay, ed., *Cinquièmes journées d’études et d’échanges sur la compression et la représentation des signaux audiovisuels (CORESA’99)*, Sophia-Antipolis, France 14–15 Jun. 1999, Centre de recherche et développement de France Télécom (Cnet), EURÉCOM, Conseil général des Alpes-Maritimes and Télécom Valley.
[10] Y. Nakaya and H. Harashima. An iterative motion estimation method using triangular patches for motion compensation. In Proc. of SPIE, Vol. 1605, pp. 546-557, 1991.
[11] N. F. Johnson, Z. Duric and S Jajodia. ‘Recovery of watermarks from distorted images’. In A. Pfitzmann, ed., *preliminary proceedings of the third international information hiding workshop*, Dresden, Germany, pp. 361–375, 29 Sep.–1 Oct. 1999.
[12] M. Kutter, ‘Watermarking resisting to translation, rotation and scaling’, in A. G. Tescher et al., eds, *Proceedings of SPIE international symposium on voice, video, and data communications – multimedia systems and applications*, vol. 3528, pp. 423–431, Boston, Massachusets, U.S.A., 2–4 Nov. 1998. The international Society for optical engineering. ISBN 0-8194-2989-9.
[13] F. Hartung, J. K. Su and Bernd Girod, ‘Spread-spectrum watermarking: Malicious attacks and counterattacks’, in P. W. Wong and E. J. Delp, eds, *Proceedings of SPIE international symposium on voice, video, and data communications – security and watermarking of multimedia contents*, vol. 3657, pp. 147–158, San Jose, California, U.S.A., 25–27 Jan. 1999. The Society for imaging science and technology (IS&T) and the international Society for optical engineering (SPIE). ISBN 0-8194-3128-1.
[14] Institut EURÉCOM Newsletter, no. 14, Dec. 1999, <http://www.eurecom.fr/>.
[15] J.-L. Dugelay, ‘Procédé de dissimulation d’informations binaires dans une image numérique’. French patent FR2775812, Institut Eurécom, 10 Sep. 1999. Available from <http://www.inpi.fr/>. Also available as WOFR9900485.
[16] F. A. P. Petitcolas and R. J. Anderson, ‘Evaluation of copyright marking systems’, in *proceedings of IEEE multimedia system’99*, vol. 1, pp. 574–579, 7–11 Jun. 1999, Florence, Italy.
[17] S. L. Brand, ‘Department of Defense trusted computer system evaluation criteria’. Tech. Rep. DoD 5200.28-STD, U.S. Department of Defense, Washington D.C., U.S.A., 26 Dec. 1985.