Semi-fragile watermarking for video surveillance applications

Marwen Hasnaoui, Mihai Mitrea

SEMI-FRAGILE WATERMARKING FOR VIDEO SURVEILLANCE APPLICATIONS

Marwen Hasnaoui, Mihai Mitrea

Institut Télécom ; Télécom SudParis, ARTEMIS Department

ABSTRACT

This paper advances SPYART, a novel semi-fragile watermarking scheme for MPEG-4 AVC protection. The authentication information, granting the method fragility, is provided by the Intra prediction mode types. This signature is embedded in the quantized error prediction of the DCT coefficients by an *m*-QIM technique, thus ensuring the method robustness. SPYART was evaluated under the framework of a videosurveillance application; the results exhibit fragility to content replacement (with an 1/81 frame and 3s spatial and temporal accuracy, respectively) and robustness against transcoding (MPEG-4 AVC compression by a factor of 4). As both the signature extraction and mark embedding take place at the MPEG-4 AVC syntax element level, the method also features low complexity.

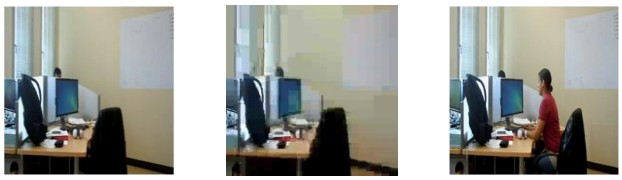
Index Terms— semi-fragile watermarking, content integrity, MPEG-4 AVC.

1. INTRODUCTION

Today, video-surveillance systems have become ubiquitous. The need to feel safe despite the high crime rate, the low cost compared to a human surveillance and the social acceptance enacted the intensive use of video-surveillance despite its intrinsic privacy intrusive character. Videosurveillance can be found today not only in particularly sensitive areas such as banks, airports and government buildings but also in public places (stadiums, parks and residential areas). For instance, more than 9400 cameras are deployed in the London public transportation [1]. This expansion of video-surveillance usage comes across with new challenges for the underlying video processing systems, as ensuring methodological support for videosurveillance content authenticity for instance. A solution can be provided by semi-fragile digital watermarking [2-5].

In its widest acceptation, a watermarking system imperceptibly (*transparently*) inserts a mark into some original content. The mark detection is performed on the watermarked content, after its processing by mundane or malicious transformations (*attacks*). A watermarking system is referred to as *robust* when the mark is always recovered, no matter the type of attack (strong compression, linear/non-linear filtering, geometrical modificatios, spatio-temporal cropping, *ect*). Robust watermarking applications are related to copyright protection [6], *in-band enrichment* [7], *etc*. A watermarking system is referred to as *fragile* when even the slightest modification of the watermarked content (brightness alteration, color changing, *ect.*) results in fails in mark detection [8]. Fragile watermarking is related to ID document authentication, bank check protection, *etc*.

A semi-fragile watermarking system should provide an application-driven trade-off between robustness and fragility. For instance, videosurveillance applications (Fig. 1.a) require robustness against mundane video processing like compression (Fig. 1.b) or file format changing but fragility against content replacement (Fig. 1.c).



(a) (b) (c)

Figure 1: Original, compressed and content changed frame.

Any real-life videosurveillance application requires a large quantity of sequences to be processed. Hence, the constraint of speed is additionally imposed and the compressed stream (*e.g.* MPEG-4 AVC) watermarking can meet it.

This paper advances a novel MPEG-4 AVC semi-fragile watermarking method jointly featuring robustness against compression and fragility to spatio-temporal cropping. The paper is structured as follows. Section 2 presents the state-of-the-art results. Section 3 is devoted to the method presentation while Section 4 validates it under the framework of the SPY ITEA2 Project [9]. Conclusions are drawn and perspectives are open in Section 5.

2. STATE-OF-THE-ART

Video authentication by means of watermarking techniques was already the object of several research studies [5, 10-12].

C. C. Wang and Y. C. Hsu [5] present a fragile watermarking algorithm to authenticate MPEG-4 AVC stream. The mark is computed as the MD5 hash function of a random generated binary sequence and inserted into the high-frequency quantized DCT coefficients of the *I* frames. While such a technique provides the ideal case of fragility and features low complexity (only the MPEG-4 AVC entropic decoding being required), it is conceptually unable to make any distinction between mundane and malicious attacks (so, it reaches the worst robustness case).

J. Zang and A. T. S. Ho [10] adopt the same principles and insert the mark in the *P* frames. The overall results are the same: a very good sensitivity to spatio-temporal alterations, low complexity but no robustness.

S. Chen and H. Leung present a semi-fragile watermarking scheme based on chaotic systems for the authentication of

individual frames in the MPEG-4 AVC stream [11]. The authentication information is represented by both the GOP index and the frame index in that GOP. This information is modulated in a chaotic signal and inserted in DCT transformed blocks of each frame by imposing local intensity relationships into a group of adjacent blocks. The insertion requires the entropic decoding, the de-quantizing and the reverse of the prediction operations, thus becoming computationally complex. Experiments carried out on a 795 frames video sequence proved robustness against JPEG compression (Q=30) and median filtering. This method also detects the temporal modifications (with frame accuracy) but the spatial modification properties were not assessed.

S. Thiemert and al [12] present a semi-fragile watermarking system devoted to the MPEG-1/2 video sequences. The marking computation and insertion is based on the properties of the entropy computed at the 8x8 block levels. The experiments are run on one sequence (whose length is not précised) encoded at 1125 kbps. The method proved both robustness (against JPEG compression with QF=50) and fragility against temporal (with 2 frame accuracy) and spatial (with a non-assessed accuracy) content changing. The main drawback of the method remains its computational complexity: the complete MPEG decoding/encoding comes across with sophisticated entropy estimation.

Tab 1 shows that the trade-off among fragility, robustness and complexity is not yet achieved. The present paper takes this challenge: it advances an MPEG-4 AVC watermarking method (further referred to as SPYART) and objectively assesses it in terms of fragility to spatio-temporal alterations, robustness to compression and complexity.

Table 1: State-of-the art synopsis.

Method/domain	Robustness	Fragility	Complexity
Wang & Hsu [6] MPEG-4 AVC	KO	Sensitive to all manipulations	Entropic decoding
Zang & Ho [7] MPEG-4 AVC	KO	Sensitive to all manipulations	Entropic decoding
Chen & Leung [8] MPEG-4 AVC	Frame-level JPEG(*QF=30*)	Temporal alterations	MPEG 4-AVC decoding
Thiemert and al.[9] MPEG-1/MPEG-2	Frame-level JPEG(*QF=50*)	Spatio-temporal alterations	MPEG 4-1/2 decoding Entropy estimation

3. SPYART METHOD

SPYART considers individual groups of k successive I frames (further referred to as *I-Group*) sampled from an MPEG-4 AVC video sequence. Within such an *I-Group*, an authentication signature is extracted from the first I frame (thus ensuring fragility) and inserted into the rest of k-1 I frames by means of a robust watermarking technique. The low complexity requirement can be met when the signature is extracted and inserted directly from/in the MPEG-4 AVC syntax elements, with minimal decoding/re-encoding.

The SPYART signature corresponds to the Intra prediction mode types in the first I frame (Section 3.1) while its insertion follows the m-QIM [13] principles, acting in the domain of the quantized 4x4 DCT coefficients of the prediction errors (Section 3.2). The mark detection and its subsequent processing in order to spatio-temporal localize the malicious alterations are presented in Section 3.3.

3.1. Signature generation

Signature generation is structured into three modules as shown in Fig 2: *feature extraction*, *binary mask generation* and *signature encoding*.

3.1.1. Feature extraction

When encoding the I frames, the MPEG-4 AVC standard can consider two types of blocks [14]: 16x16 pixel blocks for smoothed regions (corresponding to the 4 ways of achieving the $I16MB$ prediction modes) and 4x4 pixel blocks for textured areas (corresponding to the 9 ways of achieving the $I4MB$ prediction modes). *We shall further consider that the feature allowing the content authentication in a 16x16 macroblock is the size of the blocks on which the corresponding intra prediction is done.*

On the one hand, according to the MPEG coding principles, any alteration that changes the texture of the content will *a priori* change the prediction modes, thus allowing content integrity verification (related to the fragility property). On the other hand, there is no *a priori* hint about the robustness of this signature against transcoding and format changing attacks. Hence, a statistical investigation on the behavior of the block size in intra prediction under different compression attack was conducted and described in Section 4.1.

3.1.2. Binary mask generation

For the first frame from an *I-Group* (the I_0 frame in Fig 2), a binary mask is generated by assigning one bit for each macroblock $B(x,y)$ based on its extracted feature:

$$BM(x,y) = \begin{cases} 1, & if\ I4MB \\ 0, & if\ I16MB \end{cases},$$

where x and y represent positions of the B macroblock within the frame I_0 and BM is the generated binary mask.

3.1.3. Signature encoding

In order to cope with the m-QIM (multi symbol Quantizing Index Modulation) principles, the BM binary mask should be encoded into an m-arry alphabet; as in the SPYART case $m=5$, the encoding alphabet is $\{-2,-1,0,1,2\}$. For real life watermarking applications, this encoding procedure should also ensure limited error propagation: hence, a fixed-length encoding is preferred instead of an optimal (minimal average length) one. Note that according to the Shannon's first theorem, the optimal encoding average length between a binary and an m-ary alphabet is $\log_2 m$; in our case, $\log_2 5 \cong 2.31$. Hence, we considered an encoding scheme based on 3 bit overlapping blocks (with the overlap of 1 bit). The binary value of such a block gives information about the sign and the parity of the 5ary alphabet, Tab 2 and Fig 3.

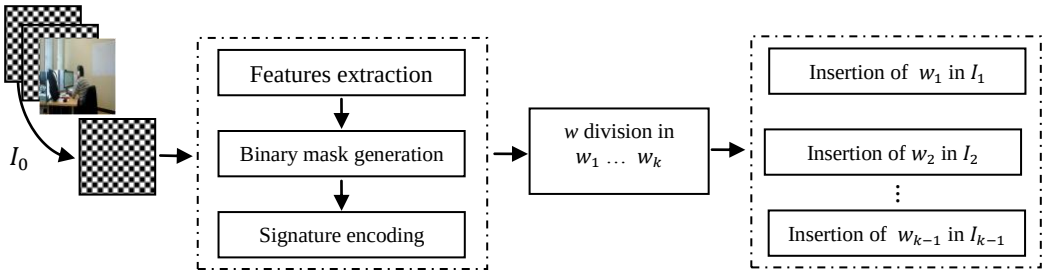


Figure 2: Mark generation $W = \{w_1, w_2, \dots, w_k\}$ based on syntax elements.

Table 2: Encoding table.

000	001	010	011	110	101	100	111
0	-1	-1	-2	1	1	2	2

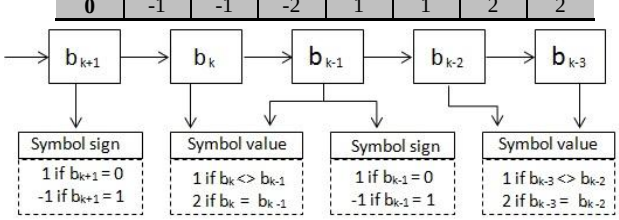


Figure 3: Signature encoding.

In order to illustrate this principle, consider the encoding of the 101001110 bit string. This bit string results into 4 overlapping groups of 3 bits: 101, 100, 011 and 110. Hence, it will be encoded as -1, -2, 2, -1; notice that both 101 and 110 are encoded as -1. The decoding procedure is applied at two levels: first, the signs of the 5ary symbol alphabet are converted into the corresponding bits and then the parity information is considered to decode the remaining sequences. In order to illustrate the decoding, reconsider the example above; we have to decode the string -1, -2, 2, -1. First, we decode the bits placed at the odd position in the bit string, by considering the signs of the 5ary symbols: 1x1x0x1x (by x we denoted the bits unknown at this stage). Further on, the encoding dictionary in Tab 2 gives the following sequence: 110100111x. In order to decode the last bit, a new symbol should be received (this is achieved by always padding the useful information with two fixed bits).

3.2. Mark embedding

For each *I-Group*, the signature generated from the first frame (I_0 in Fig 2) is shuffled (according to a private key) and divided into $k-1$ sub-marks w_i. Each sub-mark is inserted into one of the I_i frames of that *I-Group*, Fig 4. The insertion is performed within the 15 AC coefficients of 4x4 blocks of *I* frame by *m*-QIM [13] techniques.

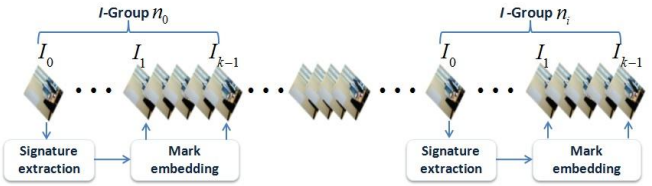


Figure 4: Mark embedding.

3.3. Mark detection and integrity verification

Consider now the watermarked and potentially corrupted video sequence, see Figs 5-6.

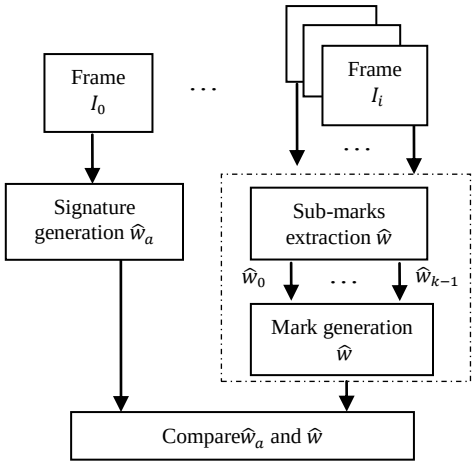


Figure 5: Integrity verification.

This sequence is first re-encoded with the original encoder parameters and then divided into *I-Groups*. The $\hat{w}_i$ sub-marks are individually extracted according to the *m*-QIM principles [13] from each $I_i, i = 1,\dots k$ frames; be there $\hat{w}$ the vector obtained by concatenating these extracted sub-marks.

In parallel, the signature corresponding to the first I_0 frame is extracted and the corresponding would-be mark w_a is computed.

As $\hat{w}$ conveys information about the original *I-Group* features and w_a about its attacked replica, by comparing these two watermarks, a decision concerning the integrity of the video content can be made. SPYART considers that an area in a frame was modified when at least 50% of the $\hat{w}$ elements extracted from that area do not match the corresponding $\hat{w}_a$ elements, see Fig 6. Consequently, SPYART has a temporal precision given by the duration of the *I-Group* and a spatial accuracy given by the size of the area on which the alteration is investigated.

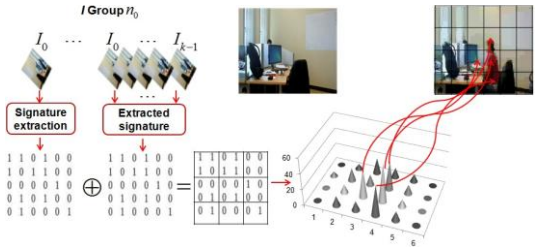


Figure 6: Spatial alterations detection.

4. FUNCTIONAL EVALUATION

The experiments were carried out on a videosurveillance corpus composed of 8 sequences of about 10 minutes each, downloaded from internet [15] or recorded under the framework of the SPY project. Their content is heterogeneous, combining city streets, highways, industrial objectives, shopping centers, *ect.*

This corpus is encoded in MPEG-4 AVC in Baseline Profile (no B frames, CAVLC entropy encoder) at 512 kbps, 576x576 pixel frames; the GOP size is set to 8.

The SPYART algorithm was applied on *I-Groups* of 3s.

Three types of experiments were performed: concerning the appropriateness of the class of the intra prediction mode for representing the authentication information inside a block (see Section 4.1), the robustness of the SPYART (see Section 4.2) and its fragility (see Section 4.3).

4.1. Intra prediction mode behavior

As already discussed in Section 3.1.1, SPYART considered as authentication information the visual content the two main classes of the MPEG-4 AVC intra prediction modes (*i.e.* the size of the block on which the intra-prediction is done – *I4MB* and *I16MB*). While the fragility properties of this information are implicitly ensured by the very MPEG-4 AVC principles, its robustness to re-encoding is further subject to investigation; in this respect, Tab 3 presents the percentage of the *I4MB* and *I16MB* blocks changed under the re-encoding attack.

Be there the whole videosurveillance corpus (the 80 min of video) encoded as described before and be there the same corpus, compressed down to 128 kbps and then re-encoded at the initial bitrate, by using the MPEG-4 AVC reference software [16]. This attack is applied according to two scenarios: (S1) the reference software is allowed to chose the re-encoded parameters and (S2) the video is re-encoded by using the initial encoding parameters.

The first two columns (corresponding to S1) in Tab 3 bring to light that the intra prediction block size is very sensitive to re-encoding: when considering these values as a noisy channel, the corresponding average error is 16%.

However, when considering the S2 case (see the last two columns in Tab 3), a better robustness is obtained, the related average error being 1.79%.

As the 95% relative errors in probability estimation were lower than $5*10^{-5}$, this experiment demonstrates that the signature we considered can meet the requirements of our targeted application, if a re-encoding with the original parameters is achieved. Actually, for practical videosurveillance applications, we can impose a fixed value for the MPEG-4 AVC QP parameter, *e.g.* QP=31, with virtually no increase in bitrate and with acceptable loss of the video visual quality (*see* Section 4.2).

Table 3: Rates of mode changes.

	S1		S2	
	I4MB	**I16MB**	**I4MB**	**I16MB**
I4MB	88.25	11.75	98.75	1.25
I16MB	20.25	79.75	2.33	97.66

4.2. Robustness

In videosurveillance context, transcoding is the most harmless authorized attack. While the Section 4.1 investigated the effects of this attack at the feature level, we are now to assess the global effectiveness of the SPYART method.

In this respect, SPYART watermarked sequence was subject to a re-encoding attack applied according to the scenario S2 described above: the attacked video is re-encoded with the initial parameters, thus ensuring the GOP re-alignments.

Further on, in order to identify the spatial content alterations, the detection procedure was applied on areas obtained by partitioning the I_0 frames with a 9x9 equidistant rectangular grid (see Fig 6). As previously explained (see Section 3.3), the decision inside each such an area is based on a majority rule, *i.e.* an area is considered as modified when more than 50% of its features are modified.

This set-up allows the robustness to be objectively assessed by the probabilities of missed detection (*i.e.* the probability of not detecting a watermark from an initially marked area), and false alarm (*i.e.* the probability of detecting a mark in an initially un-watermarked area).

Our experiments showed that the re-encoding from 512 kbps down to 128 kbps resulted in no content modification, thus demonstrating the robustness of the SPYART, with ideal values for the probabilities of missed detection and of false alarm ($P_m = 0$, $P_f = 0$).

4.3. Fragility

This section investigates the SPYART fragility (sensitivity) to content changing attacks, *i.e.* semantic manipulation of objects (such as persons or cars). The content is considered as being attacked when one object is moved, deleted or substituted. To simulate this attack, we used a piece of code that tampers the videos by changing 1/16 of the frame content arbitrarily. For each video sequence in the corpus, we applied such an attack to sequences of successive frames (between 9s and up to 3min).

In order to spatially locate alterations, we kept the same conditions as in Section 3.3: the I_0 frames in each *I-Group* are portioned in 81 areas, according to 9x9 equidistant rectangular grid. From the fragility point of view, an ideal watermarking method will fail in detecting the mark from each and every area which was subject to content alterations. While such a behavior can be also expressed in

terms of probability of missed detection and false alarm, the literature brings to light two more detailed measures, namely the precision and the recall ratios, defined as follows [17]:

$$\text{Precision} = tp/(tp + fp), \quad \text{Recall} = tp/(tp + fn),$$

where tp is the number of true positive (*i.e.* the number of content modified areas which do not allow the mark to be recovered), fp is the false positive (*i.e.* the number of content preserved areas which do not allow the mark to be recovered) and fn is the false negative number (*i.e.* the number of content modified areas which allow the mark to be detected).

The experiments exhibit $Precision = 0.81$ and $Recall = 0.92$. As these average measures are quite far from the ideal cases ($Precision = Recall = 1$), we went further in our investigation. Fig 7 illustrates the temporal detection of alterations: the abscissa corresponds to the *I-Group* index while the ordinate is set to 1 for the *I-Groups* identified by SPYART as being modified. The content attacked sub-sequences are circled in blue.

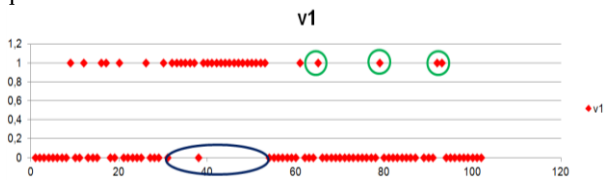


Figure 7: Alterations detection.

We can see that almost all altered *I-Groups* have been detected; however, some content-preserving *I-Groups* (circled in green) were also detected. When inspecting all the corpus on the temporal axis, it was noticed that such errors are sparse; consequently, we can introduce a post-processing decision rule: one an *I-Group* is considered as altered if at least two *I-Groups* that succeed it or precede it are detected as altered. This way, the new values for precision and recall ratios are $Precision = 0.92$ and $Recall = 0.97$. Of course, this increase in the statistical performances was obtained at the expense of decreasing the time accuracy: content modifications shorter than 9s cannot be detected.

4.4. Transparency

While the robustness-fragility trade-off is the strongest constraint for videosurveillance, the transparency constraint is somewhat less restrictive than in the case of other watermarking applications. Actually, the videosurveillance has no artistic / subjective purpose and its processing should only serve semantic relevant object management (be it by human or computer means). In order to evaluate the transparency, our study considers objective quality metrics evaluated on the same corpus. The following values were obtained: PSNR (peak signal to noise ratio) = 40 dB, PMSE (peak mean square error) = 3 10^{-3} and NCC (normalized cross-correlation) =0.99.

5. CONCLUSION

The paper presents a novel semi-fragile watermarking method for MPEG-4 AVC videosurveillance stream reaching the trade-off among robustness against transcoding, fragility to spatio-temporal cropping and computational complexity (only MPEG-4 AVC entropic decoding). In perspective, the fragility/robustness properties against other types of attacks, like special types of filtering, required by object detection and motion tracking will be studied. Establishing the proper noisy channel modeling this watermarking method and theoretically assessing the *m*-ary encoding procedure that we consider in section 3.1.3 will be also part of our future work.

6. REFERENCES

[1] J.Durand www.liberation.fr/page.php?article=315627 La caméra, nouvelle arme des policiers européens, 2005.

[2] M. Mitrea, F. Prêteux and J. Nunez, "Procédé de Tatouage d'une Séquence Vidéo," *French patent No. 05 54132 (December 2005); in the name of GET/INT et SFR*. European extension under the number 1804213 (*cf.* Bulletin européen des brevets No. 2007/27 from July 4th, 2007).

[3] M. Mitrea, T. Zaharia, F. Prêteux and A. Vlad, "Wavelet versus DCT – based spread spectrum watermarking of image databases," *Proc. SPIE*, pp. 37-46.

[4] S. Duta, M. Mitrea, F Prêteux and M. Belhaj, "MPEG-4 AVC domain watermarking transparency," *Proc. SPIE*, pp. 1-12, March 2008.

[5] C. C. Wang and Y. C. Hsu, "Fragile watermarking scheme for H.264 video stream authentication," *Optical Engineering*, pp. 49-52, February 2010.

[6] I J. Cox, M.L. Miller and J.A. Bloom, "Digital Watermarking," *Academic Press*, 2002.

[7] M. Mitrea and F. Préteux, "From Watermarking to In-Band Enrichement: Theoretical and applicative Trends," *Proc. SPIE 7248*, 2009

[8] J. Titman, A. Steinmetz and R. Steinmetz, "Content based digital signature for motion picturesauthentication and content fragile watermarking, Multimedia computing and systems," *Multimedia computing and systems IEEE International Conference on*, Italy, pp. 209-213, 1999.

[9] www.itea2.org

[10] J. Zang and A. T. S. Ho, "Efficient Video Authentication for H.264/AVC," *Proceedings of the First International Conference on Innovative Computing, Information and Control*, 2006.

[11] S. Chen and H. Leung, "Chaotic Watermarking for Video Authentication in Surveillance Applications," *IEEE Trans On Circuits and Systems For Video Technology*, pp. 704–709, Mai 2008.

[12] S. Thiemert, H. Sahbi and M. Steinebach, "Using entropy for image and video authentication watermarks," *SPIE-IS&T*, 2006.

[13] M. Hasnaoui, M. Mitrea, M. Belhaj and F. Préteux, "*m*QIM principles for MPEG-4 AVC watermarking," *IST&SPIE Electronic Imaging*, San Francisco, CA-USA, Janvier 2011.

[14] "Overview of H.264," H.264/MPEG-4 Part 10 White Paper, www.vcodex.com.

[15] K. Hühring, H.264 Reference Software Group [Online] Availabal : www.iphome.hhide, joint model 86 (JM86).

[16] www.webcamendirect.net

[17] M. Buckland, F. Gey, " The relationship between Racall and Precision ," *Jornal of the American Society for information Science* , pp. 12-19, 1994.