Abstract: - In this paper, we propose novel algorithmic models based on feature transformation in cross-modal subspace and their multimodal fusion for different types of residue features extracted from several intra-frame and inter-frame pixel sub-blocks in video sequences for detecting digital video tampering or forgery. An evaluation of proposed residue features – the noise residue features and the quantization features, their transformation in cross-modal subspace, and their multimodal fusion, for emulated copy-move tamper scenario shows a significant improvement in tamper detection accuracy as compared to single mode features without transformation in cross-modal subspace.

Key-Words: - Video tamper detection, multimodal fusion, cross-modal analysis

1 Introduction

Digital Image tampering or forgery has become major problem lately, due to ease of artificially synthesizing photographic fakes- for promoting a story by media channels and social networking websites. This is due to significant advances in computer graphics and animation technologies, and availability of low cost off-the-shelf digital image manipulation and cloning tools. With lack of proper regulatory frameworks and infrastructure for prosecution of such evolving cyber-crimes, there is an increasing dissatisfaction about increasing use of such tools for law enforcement, and a feeling of cynicism and mistrust among the civilian operating environments.

Another problem this has lead to, is a slow diffusion of otherwise extremely efficient image based surveillance and identity authentication technologies in real-world civilian operating scenarios. In this paper we propose a novel algorithmic framework for detecting image tampering and forgery based on extracting noise and quantization residue features, their transformation in cross-modal subspace and their multimodal fusion for intra-frame and inter-frame image pixel sub blocks in video sequences. The proposed algorithmic models allow detecting the tamper or forgery in low-bandwidth video (Internet streaming videos), using blind and passive tamper detection techniques and attempt to model the source signatures embedded in camera pre-processing chain. By sliding segmentation of image frames, we extract intra-frame and inter-frame pixel sub-block residue features, transform them into optimal cross-modal subspace, and perform multimodal fusion to detect evolving image tampering attacks, such as JPEG double compression, re-sampling and retouching. The promising results presented here can result in the development of digital image forensic tools, that can help investigate and solve evolving cyber crimes.

Digital image tamper detection can use either active tamper detection techniques or passive tamper detection techniques. A significant body of work, however is available on active tamper detection techniques, which involves embedding a digital watermark into the images when the images are captured. The problem with active tamper detection techniques is that not all camera manufacturers embed the watermarks, and in general, most of the customers have a dislike towards cameras which embed watermarks due to compromise in the image quality. So there is a need for passive and blind tamper detection techniques with no watermark available in the images.

Passive and blind image tamper detection is a relatively new area and recently some methods have been proposed in this area. Mainly these are of two categories [1,2,3,4]. Fridrich [4] proposed a method based on hardware aspects, using the feature extracted from photos. This feature called sensor pattern noise is due to the hardware defects in cameras, and the tamper detection technique using this method resulted in an accuracy of 83% accuracy. Chang [5] proposed a method based on camera response function (CRF), resulting in
detection accuracy of 87%, at a false acceptance rate (FAR) of 15.58%. Chen et al. [6] proposed an approach for image tamper detection based on a natural image model, effective in detecting the change of correlation between image pixels, achieving an accuracy of 82%. Gou et al [7] introduced a new set of higher order statistical features to determine if a digital image has been tampered, and reported an accuracy of 71.48%. Ng and Chang [8] proposed bi-coherence features for detecting image splicing. This method works by detecting the presence of abrupt discontinuities of the features and obtains an accuracy of 80%. Popescu and Farid [3] proposed different CFA (colour filter array) interpolation algorithms within an image, reporting an accuracy of 95.71% when using a 5x5 interpolation kernel for two different cameras. A more complex type of passive tamper detection technique, known as “copy-move tampering” was investigated by Bayram, Sencar, Dink and Memon [1,2] by using low cost digital media editing tools such as Cloning in Photoshop. This technique usually involves covering an unwanted scene in the image, by copying another scene from the same image, and pasting it onto the unwanted region. Further, the tamperer can use retouching tools, add noise, or compress the resulting image to make it look genuine and authentic. Finally, detecting tampers based on example-based texture synthesis scheme was proposed by Criminisi et al [9] that is based on filling in a region from sample textures. It is one of the state-of-the-art image inpainting or tampering schemes.

In a typical crime investigation scenario, when there is a suspicion over authenticity of the photo or video footage, the procedure followed by law enforcement agencies is to ask the photographer to turn in the camera by which the photo was taken. Then using the images captured by the camera and the images under suspicion, the camera source features (camera response function for example) get extracted, and using the statistics of the feature pattern, the two image sources are compared. However, the success of this approach relies on availability of camera source model for comparison, and establishing the possible tampering by comparison. Firstly, this is not quite a blind and passive tamper detection approach, and secondly, availability of reference model (camera source) is not possible in low-bandwidth Internet streaming scenario, where the tamperer leaves no trace of original source, and only tampered or forged video is available.

We propose a novel approach to deal with such tamper situations. The approach is based on detecting the tamper from the multiple image frames, by extracting noise and quantization residue features in intra-frame and inter-frame pixel sub blocks, transforming them into cross-modal subspace to extract the correlation properties, and establish possible tampering of video. The approach is blind and passive, and is based on the hypothesis, that a typical tampering attacks such as double compression, re-sampling and retouching can inevitably disturb the correlation properties of the pixel sub-blocks within a frame (intra-frame) as well as between the frames (inter-frame) and can distinguish the fingerprints or signatures of genuine video from tampered video frames.

The rest of the paper is organized as follows. Next Section describes the basic imaging pipeline used in digital cameras, and the source features that can leave a fingerprint on the image frames. If a tamper is attempted, the correlation distribution between intra-frame and inter-frame pixel blocks does not remain intact, giving clues about tampering. Section 3 describes the modeling of intra-frame and inter-frame features for extracting the feature correlation statistics. The proposal for multimodal fusion of the extracted features is described in Section 4. The details of the experimental results for the proposed algorithmic models are described in Section 5. The paper concludes in Section 6 with some conclusions and plan for further work.
2 Camera Processing Pipeline

The processing pipeline once the images or video is captured is shown in Fig.1. First, the camera sensor (CCD) captures the natural light passing through the optical system. Generally, in consumer digital cameras, every pixel is detected by a CCD detector, and then passed through different color filters called Color Filter Array (CFA). Then, the missing pixels in each color plane are filled in by a CFA interpolation. Finally, operations such as demosaicing, enhancement and gamma correction are applied by the camera, and converted to a user-defined format, such as RAW, TIFF, and JPEG, and stored in the memory.

3 Residue Features in Cross Modal Subspace

Different residue features described in the previous Section were first extracted from 32 x 32 pixel intra-frame and inter-frame pixel sub-blocks of the video sequences. These features were then transformed into cross-modal subspace by performing three different types of correlation processing. They are the Latent Semantic Analysis (LSA), the Cross-modal Factor Analysis (CFA), and the Canonical Correlation Analysis (CCA). The details of these subspace methods is given below:

3.1 Latent Semantic Analysis

Latent semantic analysis (LSA) is used as a powerful tool in text information retrieval to discover underlying semantic relationship between different textual units (e.g. keywords and paragraphs) [10]. It is possible to detect the semantic correlation between visual faces and its associated speech based on LSA technique. The method consists of three major steps: the construction of a joint intra-frame and inter-frame pixel sub-block feature space, the normalization, the singular value decomposition (SVD), and the semantic association measurement.

Given n inter-frame features and m inter-frame features for each of the t pixel sub-blocks of size 32 x 32 pixels, the joint feature space can be expressed as:

$$X = \{V_1, \ldots, V_1, \ldots, V_n, A_1, \ldots, A_1, \ldots A_m\}$$

(1)

where

$$V_i = (v_i(1), v_i(2), \ldots, v_i(t))^T$$

(2)

and

$$A_i = (a_i(1), a_i(2), \ldots, a_i(t))^T$$

(3)

Various intra-frame and inter-frame pixel sub-blocks can have quite different variations. Normalization of each feature according to its maximum elements (or
certain other statistical measurements) is thus needed and can be expressed as:

\[
\hat{X}_j = \frac{X_j}{\max(\frac{X_j}{abs(X_j)})} \quad \forall j
\]  

(4)

After normalization all elements in normalized matrix have values between –1 and 1. SVD (Singular Value Decomposition) can then be performed as follows:

\[
\hat{X} = S \cdot V \cdot D^T
\]  

(5)

where S and D are matrices composing of left and right singular vectors and V is diagonal matrix of singular values in descending order. Keeping only the first and most important k singular vectors in S and D, we can derive an optimal approximation with reduced feature dimensions, where semantic information between intra-frame and inter-frame pixel sub blocks is mostly preserved.

### 3.2 Cross Modal Factor Analysis

In this approach intra-frame and inter-frame pixel sub-blocks are treated as two separate subsets and we focus only on the semantic patterns between these two subsets. Under the linear correlation model, the problem then is to find the optimal transformations that can best represent the coupled patterns between the features of the two different subsets. We adopt the following optimization criterion to obtain the optimal transformations:

Given two mean centered matrices X and Y, which compose of row-by-row coupled samples from two subsets of features, we want orthogonal transformation matrices A and B that can minimize the expression:

\[
\|[X - Y]^T\|^2 = \text{trace}(XAA^T - 2X + Y) - 2\text{trace}(XAYB^T) + 2\text{trace}(YB^TX^T) - 2\text{trace}(YAB^TY) + \text{trace}(YY^T)
\]

(6)

where trace of a matrix is defined to be the sum of the diagonal elements. We can easily see from above that matrices A and B which maximize \(\text{trace}(XAB^TY)\) will minimize Eqn. 3. It can be shown that such matrices are given by:

\[
\begin{align*}
A &= S_{xy} \\
B &= D_{xy}
\end{align*}
\]

where

\[
X^TY = S_{xy} \cdot V_{xy} \cdot D_{xy}
\]  

(9)

With the optimal transformation matrices A and B, we can calculate the transformed version of X and Y as follows:

\[
\begin{align*}
\hat{X} &= X \cdot A \\
\hat{Y} &= Y \cdot B
\end{align*}
\]  

(10)

Corresponding vectors in \(\hat{X}\) and \(\hat{Y}\) are thus optimized to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset.

### 3.3 Canonical Correlation Analysis

Following the development of the previous Section, we can adopt a different optimization criterion: Instead of minimizing the projected distance, we attempt to find transformation matrices A and B that maximize the correlation between X.A and Y.B. Given two mean centered matrices X and Y as defined in the previous section, we seek matrices A and B such that

\[
correlation(XA, XB) = correlation(\hat{X}, \hat{Y}) = \text{diag}(\sigma_1, \ldots, \sigma_i, \ldots, \sigma_j)
\]

(11)

where

\[
\hat{X} = Y \cdot B,
\]

And

\[
1 \geq \sigma_1, \ldots, \sigma_i, \ldots, \geq \sigma_j \geq 0.
\]

\(\sigma_i\) represents the largest possible correlation between the \(i^{th}\) translated...
features in $\vec{X}$ and $\vec{Y}$. The CCA analysis is described in further detail in [11]. Next Section describes the multimodal fusion protocol used for combining different correlation residue features.

4 Multimodal Fusion of Residue Correlation Features

In this Section, we describe the multimodal fusion scheme used for combining different types of residue features and transformed features in cross-modal subspace (described in Section II and Section III) for intra-frame and inter-frame pixel-sub-blocks. From preliminary experimentation, we found that not all pixel sub-blocks are identically correlated. Some are highly correlated, some are loosely correlated, and some are mutually independent. So we extract different correlation components between pixel sub-blocks using different algorithms. The algorithm for extracting highly correlated components and feature fusion of these components is described now.

4.1 Feature Fusion of Highly Correlated Components

Let $f_A$ and $f_L$ represent the noise residue features based on principal component analysis of intra-frame and inter-frame pixel sub-blocks. Let $A$ and $B$ represent the correlation, transformation matrices. One can apply LSA, CCA or CFA to find two new feature sets $f'_A = A^T f_A$ and $f'_L = B^T f_L$ such that the between-class cross-modal association coefficient matrix of $f'_A$ and $f'_L$ is diagonal with maximised diagonal terms. However, maximised diagonal terms do not necessarily mean that all the diagonal terms exhibit strong cross-modal association. Hence, one can pick the maximally correlated components that are above a certain correlation threshold $\theta$. Let us denote the projection vector that corresponds to the diagonal terms larger than the threshold $\theta$ by $\vec{w}_A$ and $\vec{w}_L$. Then the corresponding projections of $f_A$ and $f_L$ are given as:

$$\vec{f}_A = \vec{w}_A^T . f_A \quad (12)$$

$$\vec{f}_L = \vec{w}_L^T . f_L \quad (13)$$

Here $\vec{f}_A$ and $\vec{f}_L$ are the correlated components with high correlation, that are embedded in $f_A$ and $f_L$. By performing feature fusion of highly correlated intra-frame and inter-frame components corresponding to noise residue features, we obtain the optimized feature fused vector in the cross-modal subspace:

$$\vec{f}_{AL} = [\vec{f}_A \quad \vec{f}_L] \quad (14)$$

4.2 Score Level Fusion of Mutually Independent Components

In the Bayesian framework, late fusion or score fusion can be performed using the product rule assuming statistically independent modalities. Various methods have been proposed in the literature [12] as an alternative to the product rule such as the max rule, the min rule and the reliability-based weighted summation rule. We can compute joint scores as a weighted summation:

$$\rho(\lambda_r) = \sum_{n=1}^{N} w_n \log P(f_n|\lambda_r) \text{ for } r = 1, 2, \ldots, R \quad (15)$$

Where $\rho_n(\lambda_r)$ is the logarithm of the class-conditional probability $P(f_n|\lambda_r)$ for the $n^{th}$ modality, with class $\lambda_r$, and $w_n$ denotes the weighting coefficient for modality $n$, such that $\sum_n w_n = 1$. Note that when $w_n = \frac{1}{N} \forall n$, Eqn. 15 is equivalent to the product rule. Since the $w_n$ values can be regarded as the reliability values of the classifiers, this combination method is also referred to as RWS (Reliability Weighted Summation) rule [12,14]. The statistical and the numerical range of these likelihood scores vary from one classifier to another. Thus using sigmoid and variance normalization as described in [14], the likelihood scores can be normalized to be within the (0, 1) interval before the fusion process. The hybrid fusion vector is finally obtained by late(score) fusion of feature fused highly correlated components ($\vec{f}_{AL}$) with correlated and mutually independent noise residue features extracted from intra-frame and inter-frame image sub-blocks with weights selected using RWS rule.
Automatic Weight Adaptation

For the RWS rule, the fusion weights are chosen empirically, whereas for the automatic weight adaptation, a mapping needs to be developed between an intra-frame reliability estimate and the modality weightings. The late fusion scores can be fused via addition or multiplication as shown in Eqn. 16 and 17. Both methods were investigated and it was found that the results achieved for both were similar (based on empirically chosen weights). However, additive fusion has been shown to be more robust to classifier errors [12, 14], and should perform better when the fusion weights are automatically, rather than empirically determined. Hence the results for additive fusion only, are presented in this paper. Prior to late fusion, all scores were normalized to fall into the range of [0, 1], using min-max normalisation.

\[ P(S_i | x_A, x_V) = \alpha P(S_i | x_A) + \beta P(S_i | x_V) \]  
\[ P(S_i | x_A, x_V) = P(S_i | x_A)^{\alpha} \times P(S_i | x_V)^{\beta} \]

where:

\[ \alpha = \begin{cases} 0 & c \leq 1 \\ 1 + c & -1 < c < 0 \\ 1 & c \geq 0 \end{cases} \]

\[ \beta = \begin{cases} 1 & c \leq 0 \\ 1 - c & 0 < c < 1 \\ 0 & c \geq 1 \end{cases} \]

where \( x_A \) and \( x_V \) refer to the audio test utterance and visual test sequence/image respectively.

To carry out automatic fusion, that adapts to varying noise conditions, a single parameter \( c \), the fusion parameter, is used to define the weightings; the intra-frame pixel sub-block weight \( \alpha \) and the inter-frame pixel sub-block weight \( \beta \), i.e., both \( \alpha \) and \( \beta \) dependent on \( c \). Fig. 4 and Eqn. 17 show how the fusion weights, \( \alpha \) and \( \beta \), depend on the fusion parameter \( c \). Higher values of \( c \) (>1) place more emphasis on the intra-frame module whereas lower values (<0) place more emphasis on the inter-frame module. For \( c \geq 1 \), \( \alpha = 1 \) and \( \beta = 0 \), hence the fused decision is based entirely on the intra-frame pixel sub block likelihood score, whereas, for \( c \leq -1 \), \( \alpha = 0 \) and \( \beta = 1 \), the decision is based entirely on the inter-frame pixel sub-block likelihood score. So in order to account for varying noise conditions, only \( c \) has to be adapted.

The reliability measure was the intra-frame likelihood score \( \rho_n(\lambda,\tau) \). As the intra-frame SNR decreases, this reliability measure decreases in absolute, and becomes closer to threshold for genuine image likelihoods. Under clean test conditions, this reliability measure increases in absolute value because the genuine image model yields a more distinct score. So, a mapping between \( \rho \) and \( c \) can automatically vary \( \alpha \) and \( \beta \) and hence place more/less emphasis on the intra-frame scores. To determine the mapping function \( c(\rho) \), the values of \( c \) which provided for optimum fusion, \( c_{\text{opt}} \), were found by exhaustive search for the N tests at each SNR levels. This was done by varying \( c \) from -1 to +1, in steps of 0.01, in order to find out which \( c \) value yielded the best performance. The corresponding average reliability measures were calculated, \( \rho_{\text{mean}} \), across the N test utterances at each SNR level.

\[ c(\rho) = c_{\text{opt}} + \frac{h}{\exp[d(\rho + \rho_{\text{os}})]} \]

A sigmoid function was employed to provide a mapping between the \( c_{\text{opt}} \) and the \( \rho_{\text{mean}} \) values, where \( c_{\text{os}} \) and \( \rho_{\text{os}} \) represent the offsets of the fusion parameter and reliability estimate respectively; \( h \) captures the range of the fusion parameter; and \( d \) determines the steepness of the sigmoid curve.

Experimental Results

The video sequence data from Internet streamed movies was collected and partitioned into separate
subsets based on different actions and genres. Figure 5 shows screenshots corresponding to different actions, along with emulation of copy move tampered scenes and the detection of tampered regions with the proposed approach.

![Fig. 5: Row 1: Screenshots from Internet streamed video sequences; Row 2: Copy-move tamper emulation for the scene; Row 3: Detection of tampered regions in the scene](image)

**TABLE 1: EVALUATION OF NOISE RESIDUE FEATURES FOR EMULATED COPY-MOVE TAMPER ATTACK (% ACCURACY)**; $\tilde{f}_{\text{Intra}}$ (RESIDUE FEATURES WITH CROSS-MODAL TRANSFORMATION); $f_{\text{Intra}}$ (RESIDUE FEATURES WITHOUT CROSS-MODAL TRANSFORMATION)

<table>
<thead>
<tr>
<th>Internet streamed movie data subset</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue Features in Cross-Modal Subspace</td>
<td>CFA</td>
</tr>
<tr>
<td>$f_{\text{Intra}}$</td>
<td>85.2</td>
</tr>
<tr>
<td>$f_{\text{Inter}}$</td>
<td>83.8</td>
</tr>
<tr>
<td>$f_{\text{Intra}}-\text{Inter}$</td>
<td>83.8</td>
</tr>
<tr>
<td>$\tilde{f}_{\text{Intra}}$</td>
<td>82.18</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>89.7</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>90.68</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Inter}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>90.26</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Inter}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>92.06</td>
</tr>
<tr>
<td>$\tilde{f}_{\text{Intra}}$</td>
<td>92.06</td>
</tr>
</tbody>
</table>

Different sets of experiments were conducted to evaluate the performance of the proposed residue features in correlation sub-space and their fusion in terms of tamper detection accuracy. The experiments involved a training phase and a test phase. In the training phase a Gaussian Mixture Model for each video sequence from data base was constructed. In the test phase, copy-move tamper attack was emulated by artificially tampering the training data.

The tampered processing involved copy cut paste of small regions in the images and hard to view affine artefacts. Two different types of tampers were examined. An intra-frame tamper, where the tampering occurs in some of the pixel sub-blocks within the same frame, and inter-frame tamper, where pixel sub-blocks from adjacent frames were used. However, in this paper, we present and discuss results for the intra-frame tamper scenario only. Figure 5 shows some sample results for intra-frame tamper scenario. As can be seen from Table 1 and Table 2, which show the tamper detection results in terms of % accuracy, the performance of ordinary features fusion of both noise residue and quantization residue features can be enhanced by feature transformation in cross-modal subspace and their multimodal fusion.

**TABLE 2: (% ACCURACY) PERFORMANCE FOR NOISE AND QUANTIZATION RESIDUE FEATURES FOR BEST PERFORMING FEATURES IN CROSS-MODAL SUBSPACE**

<table>
<thead>
<tr>
<th>% Accuracy</th>
<th>Noise Residue Features</th>
<th>Quantization Residue Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different Residue features and their fusion</td>
<td>CFA features</td>
<td>CFA features</td>
</tr>
<tr>
<td>$f_{\text{Intra}}$</td>
<td>85.2</td>
<td>84.3</td>
</tr>
<tr>
<td>$f_{\text{Inter}}$</td>
<td>83.8</td>
<td>82.36</td>
</tr>
<tr>
<td>$f_{\text{Intra}}-\text{Inter}$</td>
<td>83.8</td>
<td>81.1</td>
</tr>
<tr>
<td>$\tilde{f}_{\text{Intra}}$</td>
<td>82.18</td>
<td>84.19</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>89.7</td>
<td>89.7</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Inter}} + f_{\text{Intra}}-\text{Inter}$</td>
<td>90.26</td>
<td>90.7</td>
</tr>
<tr>
<td>$f_{\text{Inter}}$</td>
<td>90.26</td>
<td>89.79</td>
</tr>
<tr>
<td>$f_{\text{Intra}} + f_{\text{Inter}}$</td>
<td>90.26</td>
<td>89.46</td>
</tr>
<tr>
<td>$f_{\text{Inter}}-\text{Inter}$</td>
<td>92.06</td>
<td>90.23</td>
</tr>
</tbody>
</table>

For the feature fusion of the highly correlated components, the accuracy improves from 84.3% to
85.2% for CFA analysis for noise residue features. Since each frame also carries mutually independent information in pixel sub-blocks, the overall performance can be enhanced with hybrid fusion, with an optimal combination of the feature-level and the score level fusion of feature vectors from intra-frame, inter-frame and transformed intra and inter-frame pixel sub-blocks in cross-modal subspace.

For the feature fusion of the highly correlated components, the accuracy improves from 84.3% to 85.2% for CFA analysis for noise residue features. Since each frame also carries mutually independent information in pixel sub-blocks, the overall performance can be enhanced with hybrid fusion, with an optimal combination of the feature-level and the score level fusion of feature vectors from intra-frame, inter-frame and transformed intra and inter-frame pixel sub-blocks in cross-modal subspace.

Also, with the noise residue features, the hybrid fusion involving late fusion of intra-frame features with feature-level fusion of highly correlated intra and inter-frame features based on CFA analysis yields a best accuracy of 92.06%. Similar improvement in tamper detection accuracy was observed for different combinations of highly correlated component and independent component fusion for the quantization residue features. For both feature sets, around 22% improvement in accuracy was achieved with inclusion of highly correlated components (CMA-transformed) features, and the subsequent multimodal fusion as compared to use of uncorrelated component fusion. It can also be noted that all the multimodal hybrid fusion modes (last four rows in Table 1 and last 2 rows in Table 2) resulted in synergistic fusion, with the % accuracy better than baseline intra-frame only and inter-frame only accuracies of 83.8% and 85.2% for noise residue features and 82.86% and 84.3% for the quantization residue features.

7 Conclusions

In this paper, we investigated a novel approach for video tamper detection in low-bandwidth Internet streamed videos using residue features from intra-frame and inter-frame pixel sub-blocks, their transformation in cross-modal subspace and the subsequent multimodal fusion. The evaluation of two different residue features, the noise and the quantization residue features for emulated copy-move tamper scenario show the potential of proposed blind and passive tamper detection approach for applications where the establishing the identity of the camera source is not available. The feature transformation of residue features in cross-modal subspace and their subsequent multimodal fusion of intra-frame and inter-frame features models the camera source signatures and allows blind and passive tamper detection. An accuracy of around 92% was achieved for hybrid fusion with residue features transformed in cross-modal subspace, an improvement of around 22% compared to fusion without transformation in the cross-modal subspace. The performance for quantization residue features for all the experiments was quite close to noise residue features. Further work involves modeling and feature extraction of other source signatures from image sequences and testing with low bandwidth Internet streamed video sequence with multiple tamper attacks.

References:


