2010 Fourth International Conference on Network and System Security

Digital Video Tamper Detection Based on Multimodal fusion of Residue Features

Girija Chetty *Member, IEEE,* Monica Biswas and Rashmi Singh

***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.**

I. 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 the 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 novel and 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.

Manuscript received 16th May, 2010. The first author is a member of IEEE and is with Faculty of Information Sciences and Engineering, University of Canberra, and the second and third authors are with Video Analytics Pty. Ltd. The contact e-mail for the corresponding author is girija.chetty@canberra.edu.au

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 embedded 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

978-0-7695-4159-4/10 $26.00 © 2010 IEEE
DOI 10.1109/NSS.2010.8

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 impainting 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.

II. 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 colour filters called Color Filter Array (CFA). Then CFA interpolation is used to fill in the missing pixels. 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.

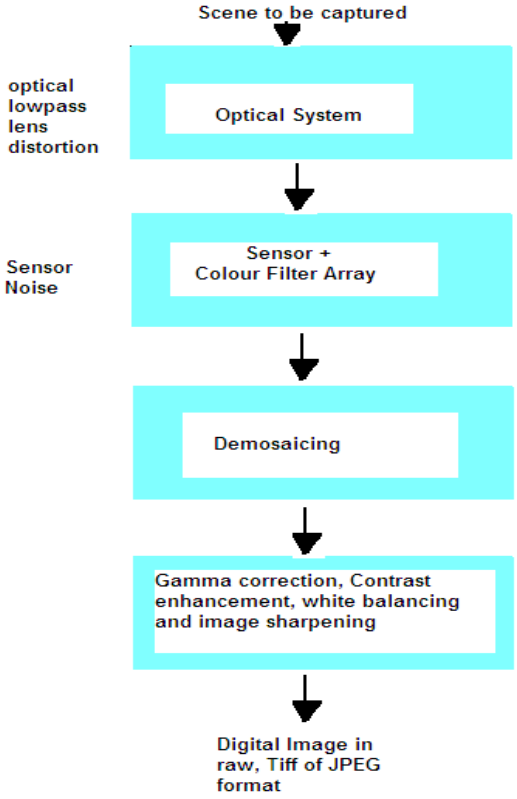

Fig 1: Camera processing pipeline

Since the knowledge about the source and exact processing (details of the camera) used is not available for application scenarios considered in this work (low-bandwidth Internet streaming video), which may not be authentic and already tampered, we extract a set of residual features for pixel sub-blocks within the frame and between adjacent frames from the video sequences. These residual features try to model and extract the fingerprints for source level processing within any camera, such as denoising, quantization, compression, contrast enhancement, white balancing, image sharpening etc. In this work, we use only two types of residual features: noise residue features and quantization residue features. An example how noise residue features can be extracted from intra-frame and inter-frame pixel sub-

blocks is shown in Figure 2 and Figure 3 below.

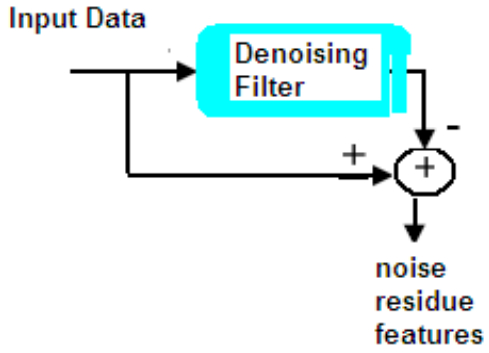

Fig 2: Extraction of noise residue features

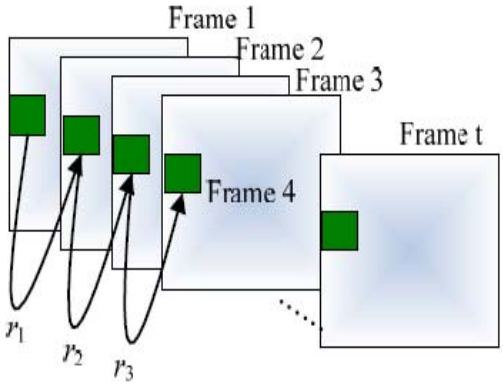

Fig 3: Extraction of intra-frame and inter-frame pixel sub-block noise residue features

In the first step, the noise residue of each video frame is extracted by subtracting the original frame from its noise-free version over a sliding window pixel sub-block. The wavelet denoising filter proposed in [13] is used to obtain the noise-free image.

In the second step, the inter-frame noise residue features are obtained by partitioning each video frame into non-overlapping blocks of size N × N. The correlation of the noise residue between the same spatially indexed blocks (inter-frame blocks) of two consecutive frames is then computed as illustrated in Fig 3.

Similar approach is used for extracting inter-frame and intra-frame pixel sub-block features corresponding to quantization residue features. A Gaussian Mixture Model (GMM) is trained with above mentioned residue features for different video sequences. To test the tamper detection ability, we emulated a copy-move tamper attack by re-painting some of the pixel sub-blocks within same frames. The correlation relationship between the intra-frame and inter-frame pixel sub-blocks was extracted by transforming the residue features in cross-modal (correlation) sub-space. The feature transformation in the cross-modal space allows detecting the variation in correlation properties between different pixel sub-blocks and localizes the tamper zones. Further, to address the situation that images might have undergone multiple tamper processes to make it look authentic and genuine, a multimodal fusion of intra-frame and inter-frame residual features in cross-modal subspace was performed. The details of extracting residue features in cross-modal sub-space and their multimodal fusion is described in next two Sections.

III. 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:

A. Latent Semantic Analysis

The Latent semantic analysis (LSA) technique is more popular in text information retrieval area, and is used to discover underlying semantic relationship between different textual units (.e.g. keywords and paragraphs) [10]. It is possible to detect the semantic correlation between inter-frame and intra-frame pixel sub-blocks using LSA technique. The analysis in this method comprises 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:

$$X = [V_1, \ldots, V_i, \ldots, V_n, A_1, \ldots, A_i, \ldots A_m], \text{ where} \quad (1)$$

$$V_i = (v_i(1), v_i(2), \ldots, v_i(t))^T \text{ and} \quad (2)$$

$$A_i = (a_i(1), a_i(2), \ldots, a_i(t))^T \quad (3)$$

Various intra-frame and inter-frame pixel sub-blocks can have quite different variations. By normalizing each feature according to its maximum elements (or certain other statistical measurements), the features can be expressed as:

$$\hat{X}_{ij} = \frac{X_{ij}}{\max(abs(X_{ij}))} \quad \forall j \quad (4)$$

All the elements in normalized matrix have values

between –1 and 1 after normalization, and the SVD (Singular Value Decomposition) can then be performed as follows:

$$\hat{X} = S \,.\, V \,.\, D^{T} \tag{5}$$

where S and D are matrices composing of left and right singular vectors and V is diagonal matrix of singular values in descending order. It is possible to derive an optimal approximation of $\hat{X}$ with reduced feature dimensions, by keeping only the first and most important *k* singular vectors in S and D, and thus the semantic information between intra-frame and inter-frame pixel sub blocks can be mostly preserved.

B. Cross-Modal Factor Analysis

In this approach intra-frame and inter-frame pixel sub-blocks are treated as two separate subsets, and under the linear correlation model, the problem is to find the optimal transformations that can best represent the correlated patterns between the features of the two different subsets.
One can use the following optimization criteria for obtaining the optimal transformations for the CFA technique: Assuming two subsets of features have been used for constructing two mean-centered matrices X and Y, orthogonal transformation matrices A and B that can minimise the expression can be shown as:

$$\| XA - YB \|_F^2 \tag{6}$$

where $\| M \|_F$ denotes the Frobenius norm of the matrix M and can be expressed as:

$$\| M \|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2} \tag{7}$$

The matrices *A* and *B* in Equation (1) define two orthogonal sub spaces where coupled data in *X* and *Y* can be projected as close to each other as possible.

$$\begin{aligned} \| XA - YB \|_F^2 &= trace\left((XA - YB) \,.\, (XA - YB)^{T} \right) \\ &= trace\left(XAA^{T} X^{T} + YBB^{T} Y^{T} - XAB^{T} Y^{T} - YBA^{T} X^{T} \right) \\ &= trace\left(XX^{T} \right) + trace\left(YY^{T} \right) - 2 \,.\, trace\left(XAB^{T} Y^{T} \right) \end{aligned} \tag{8}$$

where the trace of a matrix can be expressed as the sum of the diagonal elements. It can be observed that matrices A and B which maximise trace ($XAB^{T}Y^{T}$) will minimise the equation above. We can show that such matrices are represented by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases}$$

Where

$$X^{T} Y = S_{xy} \,.\, V_{xy} \,.\, D_{xy} \tag{9}$$

Once the optimal transformation matrices A and B are determined as in Equation (4), the transformed version of X and Y can be calculated as follows:

$$\begin{cases} \tilde{X} = X \,.\, A \\ \tilde{Y} = Y \,.\, B \end{cases} \tag{10}$$

The coupled relationships between the two feature subsets can be represented by corresponding vectors in $\tilde{X}$ and $\tilde{Y}$. One can find the first and most important *k* corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ using conventional Pearson correlation or mutual information calculation [15], facilitating the principal coupled patterns in much lower dimensions to be preserved. The CFA technique thus provides two advantages: reduction in feature dimension, as well as feature selection capability.

C. Canonical Correlation Analysis

A different optimization technique is used for Canonical Correlation Analysis (CCA) method. For the CCA method, the transformation matrices A and B are obtained by maximising the correlation between XA and XB, instead of minimizing the projected distance Following mathematic formulation can be used to describe this technique.

The two matrices A and B can be obtained from two mean centred matrices X and Y such that:

$$\begin{aligned} &correlation(XA, XB) = correlation(\tilde{X}, \tilde{Y}) \\ &= diag(\lambda_1, \ldots \lambda_2 \cdots \lambda_i) \end{aligned} \tag{11}$$

where $\tilde{X} = Y \,.\, B$, and $1 \geq \sigma_1 \geq, \cdots, \sigma_i, \cdots, \geq \sigma_l \geq 0$. σ_i represents the largest possible correlation between the i^{th} translated features in $\tilde{X}$ and $\tilde{Y}$. The CCA analysis is described in further detail in [11]. Next Section describes the multimodal fusion protocol used for combining different correlation residue features.

IV. MULTIMODAL FUSION OF RESIDUE CORRELATION FEATURES

In this Section, we describe the multimodal fusion protocol 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.

A. 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, not all the diagonal terms exhibit strong cross-modal association. Hence, we can pick the maximally correlated components that are above a certain correlation threshold θ. If we denote the projection vector that corresponds to the diagonal terms larger than the threshold θ by $\widetilde{w}_A$ and $\widetilde{w}_L$. Then the corresponding projections of f_A and f_L are given as:

$$\widetilde{f}_A = \widetilde{w}_A^T . f_A \tag{12}$$

$$\widetilde{f}_A = \widetilde{w}_L^T . f_A \tag{13}$$

Here $\widetilde{f}_A$ and $\widetilde{f}_L$ 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:

$$\widetilde{f}_{AL} = \left[\widetilde{f}_A \quad \widetilde{f}_L\right] \tag{14}$$

B. Score (Level) Fusion of Mutually Independent Components

Assuming statistically independent modalities, late fusion or score fusion can be performed using the product rule. Several other methods have been proposed in the literature on Bayesian fusion [12] as options to product rule, including the max rule, the min rule and the RWS reliability-based weighted summation rule. We can compute joint scores as a weighted summation:

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

The Eqn. 15 is equivalent to product rule, with $\rho_n(\lambda_r)$ as the logarithm of the class-conditional probability $P(f_n|\lambda_r)$ for the n^{th} modality, with class λ_r, and w_n denoting the weighting coefficient for modality n, such that $\Sigma_n w_n = 1$. Note that when $w_n = \frac{1}{N} \quad \forall n$. This fusion protocol can also be described as RWS (Reliability Weighted Summation) rule [12,14], since the w_n values can be regarded as the reliability values of the classifiers. There could be significant variation from one classifier to another in terms of statistical and numerical range. By using sigmoid and variance normalization [14], the likelihood scores can be normalized to be within the (0, 1) interval before the fusion process. The composite fusion vector is finally obtained by late(score) fusion of feature fused highly correlated components ($\widetilde{f}_{AL}$) with correlated and mutually independent noise residue features extracted from intra-frame and inter-frame image sub-blocks with weights selected using RWS rule.

V. AUTOMATIC WEIGHT ADAPTATION

We investigated an automatic weight adaptation technique in addition to RWS rule (where the fusion weights are chosen empirically). For automatic weight adaptation, a mapping was developed between an intra-frame reliability estimate and the modality weightings. As shown in Eqn. 16 and 17, the late fusion scores can be fused via addition or multiplication. The additive fusion technique has been shown to be more robust to classifier errors [12, 14], and should perform better when the fusion weights are determined automatically, rather than on an empirical basis. Prior to late fusion, all scores were normalized to fall into the range of [0, 1], using min-max normalisation.

$$P(S_i|x_A, x_v) = \alpha P(S_i|x_A) + \beta P(S_i|x_v) \quad (16)$$

$$P(S_i|x_A, x_v) = P(S_i|x_A)^{\alpha} \times P(S_i|x_v)^{\beta} \quad (17)$$

where:

$$\alpha = \begin{Bmatrix} 0 & c \le 1 \\ 1+c & -1 < c < 0 \\ 1 & c \ge 0 \end{Bmatrix}$$

$$\beta = \begin{Bmatrix} 1 & c \le 0 \\ 1-c & 0 < c < 1 \\ 0 & c \ge 1 \end{Bmatrix}$$

where x_A and x_V refer to the intra-frame pixel sub block test utterance and inter-frame pixel sub block test sequence respectively.

For 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 α and the inter-frame pixel sub-block weight β, i.e., both α and β dependent on c. Fig. 4 and Eqn. 17 show how the fusion weights, α and β, depend on the fusion parameter c. Higher values of c (>0) place more emphasis on the intra-frame module whereas lower values (<0) place more emphasis on the inter-frame module. For $c \ge 1$, $\alpha = 1$ and $\beta = 0$, hence the fused decision is based entirely on the intra-frame pixel sub block likelihood score, whereas, for $c \le -1$, $\alpha = 0$ and $\beta = 1$, the decision is based entirely on the inter-frame pixel sub-block likelihood score. So by adapting c varying noise conditions can be accounted.

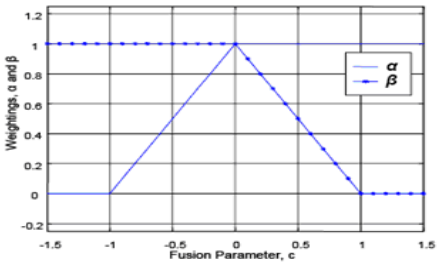

Fig. 4: The module weightings versus the fusion parameter "c"

The intra-frame likelihood score $\rho_n(\lambda_r)$ was used as a reliability measure in our study. As the intra-frame SNR decreases, the absolute value of reliability measure decreases, and becomes closer to threshold for likelihoods corresponding to genuine images in the test phase. 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 ρ and c can automatically vary α and β and hence place more/less emphasis on the intra-frame scores. The mapping function c(ρ) was obtained, and the values of c which provided for optimum fusion, c_{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, ρ_{mean}, across the N test utterances at each SNR level. Figure 4 shows the module weightings versus the fusion parameter "c".

$$c(\rho) = c_{os} + \frac{h}{\exp[d.(\rho + \rho_{os})]} \quad (18)$$

A sigmoid function was employed to provide a mapping between the c_{opt} and the ρ_{mean} values, where c_{os} and ρ_{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.

VI. 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

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

Internet streamed movie data subset	**% Accuracy**		
Residue Features in Cross-Modal Subspace	**CFA**	**CCA**	**LSA**
f_{Intra}	85.2	85.2	85.2
f_{Inter}	83.8	83.8	83.8
$f_{Intra-Inter}$	83.8	82.13	77.53
$\widetilde{f}_{Intra-Inter}$	82.18	84.82	81.91
$f_{Intra} + f_{Intra-Inter}$	89.7	89.7	89.7
$f_{Intra} + f_{Intra-Inter}$	90.68	90.86	89.29
$f_{Intra} + f_{Inter} + f_{Intra-Inter}$	90.26	90.26	90.26
$f_{Intra} + f_{Inter} + f_{Intra-Inter} + \tilde{f}_{Intra-Inter}$	92.06	91.85	91.22

For the feature fusion of the highly correlated components $\widetilde{f}_{\text{int } ra-\text{int } erp}$, 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 composite 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 $\widetilde{f}_{\text{int } ra-\text{int } erp}$, 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.

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

% Accuracy	**Noise Residue Features**	**Quantization Residue Features**
Different Residue features and their fusion	CFA features	CFA features
f_{Intra}	85.2	84.3
f_{Inter}	83.8	82.36
$f_{Intra-Inter}$	83.8	81.1
$\widetilde{f}_{Intra-Inter}$	82.18	84.19
$f_{Intra} + f_{Intra-Inter}$	89.7	88.12
$f_{Intra} + f_{Intra-Inter}$	90.68	89.79
$f_{Intra} + f_{Inter} + f_{Intra-Inter}$	90.26	89.46
$f_{Intra} + f_{Inter} + f_{Intra-Inter} + \widetilde{f}_{Intra-Inter}$	92.06	90.23

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 composite 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.

VII. CONCLUSIONS

In this paper, we present results of an investigation on 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 multimodal fusion of 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 modelling and feature extraction of other source signatures from image sequences and testing with low bandwidth Internet streamed video sequence with multiple tamper attacks.

REFERENCES

[1] S. Bayram, H. T. Sencar, and N. Memon, An Efficient and Robust Method For Detecting Copy-Move Forgery. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, Taipei Taiwan, June 2009.

[2] A. E. Dirik and N. Memon, Image Tamper Detection Based on Demosaicing Artifacts. IEEE ICIP 09, November 2009, Cairo Egypt.

[3] Alin C .Popescu and Hany Farid, "Exposing Digital Forgeries by Detecting Traces of Re-sampling",IEEE Transactions on signal processing ,Vol. 53,No.2,February 2005 .

[4] Jessica Fridrich, David Sukal and Jan Lukas, "Detection of Copy-Move Forgery in Digital Images", http://www.ws.binghamton.edu/fridrich/Research/copymove.pdf

[5] Y. F. Hsu and S. –F. Chang. "Detecting Image Splicing Using Geometry Invariants and Camera Characteristics Consistency", In ICME, Toronto, Canada, July 2006.

[6] Y. Q. Shi, C. Chen, and W. Chen, "A natural image model approach to splicing detection," in Proc. ACM Multimedia Security Workshop, pp. 51-62, Sept. 2007, Dallas, Texas.

[7] H. Gou, A. Swaminathan, and M. Wu, "Noise Features for Image Tampering Detection and Steganalysis," Proc. of IEEE Int. Conf. On Image Processing (ICIP'07), San Antonio, TX, Sept. 2007.

[8] T. T. Ng, S. –F. Chang, C. –Y. Lin, and Q. Sun, "Passive-blind Image Forensics", In Multimedia Security Technologies for Digital Rights, W. Zeng, H. Yu, and C. –Y. Lin (eds.), Elsvier, 2006.

[9] A. Criminisi, P Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," IEEE Trans. Image Process., vol.13, no.9, pp. 1200-1212, Sept. 2004

[10] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[11] M. Borga and H. Knutsson, "Finding Efficient Nonlinear Visual Operators using Canonical Correlation Analysis, " in Proc. of SSAB-2000, Halmstad, pp. 13-16.

[12] Sanderson, C. and K.K. Paliwal , "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 2409-2419, 2003.

[13] M. K. Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 6, pp. 3253–3256, Mar. 1999, Phoenix, AZ

[14] Y.Sun, Y.Shi, F.Chen, V.Chung, "Skipping Spare Information in Multimodal Inputs during Multimodal Input Fusion", Proceeding of the 2009 International Conference on Intelligent User Interfaces, IUI2009, Sanibel Island, Florida, USA, 8-11 February 2009.