Video Jitter Analysis for Automatic Bootleg Detection

Marco Visentini-Scarzanella [1], Pier Luigi Dragotti [2]

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London
SW7 2AZ London, United Kingdom
[1] marcovs@imperial.ac.uk
[2] p.dragotti@imperial.ac.uk

***Abstract*—This paper presents a novel technique for the automatic detection of recaptured videos with applications to video forensics. The proposed technique uses scene jitter as a cue for classification: when recapturing planar surfaces approximately parallel to the imaging plane, any added motion due to jitter will result in approximately uniform high-frequency 2D motion fields. The inter-frame motion trajectories are retrieved with feature tracking techniques, while local and global feature motion are decoupled through a 2-level wavelet decomposition. A normalised cross-correlation matrix is then populated with the similarities between the high-frequency components of the tracked features' trajectories. The correlation distribution is then compared with trained models for classification. Experiments with original and recaptured standard datasets show the validity of the proposed technique.**

I. Introduction

With the increasing availability of small, inexpensive video recording devices, casual movie making is now within everyone's reach. Smartphones with integrated video recording facilities are now commonplace, and with resolutions exceeding the 10 megapixels barrier and fast direct internet connections, videos can be uploaded to the internet seconds after they are captured. As a result, over one hour of video material is uploaded on websites such as YouTube every second, with traffic from mobile devices being trebled in 2011 alone [1].

This represents a great leap forward for budding directors; however it poses a significant threat in terms of copyrighted video material that can be easily covertly recorded and distributed. The recent high-profile case of the file sharing website MegaUpload being seized by the U.S. Justice Department is a case in point for the magnitude of the illegal distribution of copyrighted material.

Illegitimate video material circulating on the net may not be limited to the case of recaptured copyrighted footage. Under this particular scenario, proprietary videos aired, for example, in a cinema, are covertly recaptured with a portable camcorder and uploaded on pirate websites for download. Another possibility is to recapture a fraudulent or otherwise doctored video in order to conceal traces of forgery, claiming its authenticity.

Using recapture as an antiforensic technique covering the weak footprints left by doctoring is an attack that has been recently explored in the case of images in several works, both in the case of recapture from printouts [2], [3] and directly from LCD screens [4], [5]. Moreover, when recapturing videos the effectiveness of the attack is amplified, as the more aggressive compression strategies normally employed eliminate forgery footprints to a much greater extent.

In this paper, we propose a simple yet effective method to automatically detect recaptured videos. The proposed technique is based on the detection of high-frequency motion uniformity present in the video sequence, introduced by camera jitter during recapture. Since the method is based on characteristics at a higher level of abstraction (i.e. trajectory of image features), it is robust to antiforensic methods and poor quality input signals. The proposed method is validated with standard video sequences with different input resolutions.

The paper is organised as follows: in the next section, an overview of related work in recaptured video detection is given. Then, the proposed method is outlined with a description of its constituent stages. Finally, the method is tested on standard original and recaptured videos.

II. Related Work

Due to a greater focus on image forensics, the literature concerned with video forensics is comparatively small. Moreover, some of the methods proposed are a direct transposition to video of techniques applied to image forensics; examples include camera PRNU extraction from video frames for video copy detection, both on high quality [6] and low resolution compressed YouTube videos [7].

The approach in [8] is more exclusively specific to videos, as it aims to detect video forgeries by examining the motion regularity between interlaced fields of neighbouring frames. In [9], forged areas are identified by modelling a per-pixel noise function dependent on the image irradiance and detecting outlying image portions.

The methods above, however, are not tailored to multiple recapture scenarios, as the feature considered are either sup-

MMSP'12, September 17-19, 2012, Banff, Canada.
???-?-????-????-?/12/$??.??

pressed during the recapture process or not sensitive to its footprints. Other techniques have been devoted explicitly to the detection of recaptured cinema videos: a geometric approach is presented in [10], where reprojected videos are identified from non-zero skew parameters being introduced within the camera intrinsic matrix. However, the technique assumes a skew factor strictly equal to zero in the original video. Moreover, its robustness quickly decreases as the recapturing camera plane approaches a configuration parallel to the cinema screen. Validation on real sequences is also limited, and performed on a single 42 frame segment.

In [11], recaptures are detected by the presence of combing artifacts caused by interlaced scanning on TV screens and interlaced recordings, therefore ineffective with more modern progressive devices. The problem of automatically detecting pirated movies captured in a cinema is also examined in [12], however it requires robust watermarking of the projected video [13].

Conversely, we propose an approach explicitly devoted to the detection of recaptured videos, without assumptions on the device employed or on the video visualised. While the proposed technique is unable to provide any information on whether a video has been subjected to tampering, recapture is often a telltale sign of antiforensic activity.

III. System Workflow

The intuition behind the proposed method is shown in Fig. 1. In the most common scenario of copyrighted video recapture, the pirate obtains a bootleg recording through the covert use of a portable camcorder to capture the aired video. The video contains natural scenes with both static and independently moving elements, each with its own local trajectory. However, during recapture a global jitter noise is introduced.

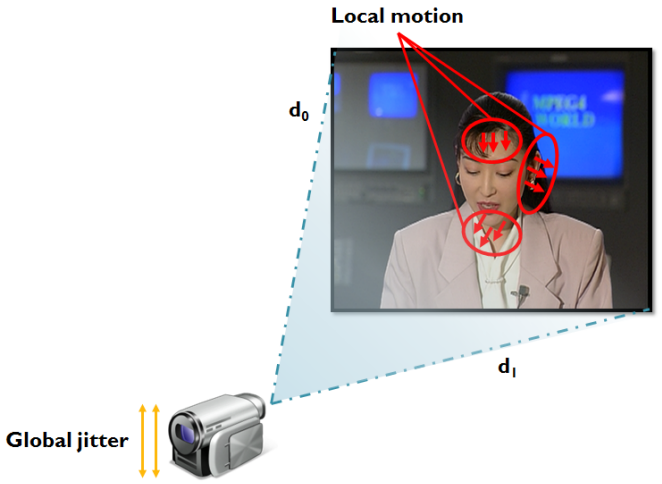


Fig. 1: Problem overview. A stable video containing both static and independently moving regions is filmed with a handheld camcorder. The jitter introduced results in uniform additive jitter noise affecting the trajectory of image features.

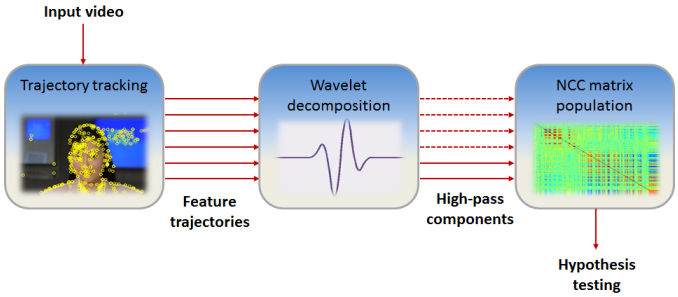


Fig. 2: System workflow. Visual features are tracked over time in the input video. The resulting trajectories are then filtered to isolate their high frequency components. Correlation between all pairs of filtered trajectories is computed to quantify jitter uniformity. Histograms of correlation and motion distributions are tested against trained distributions for classification.

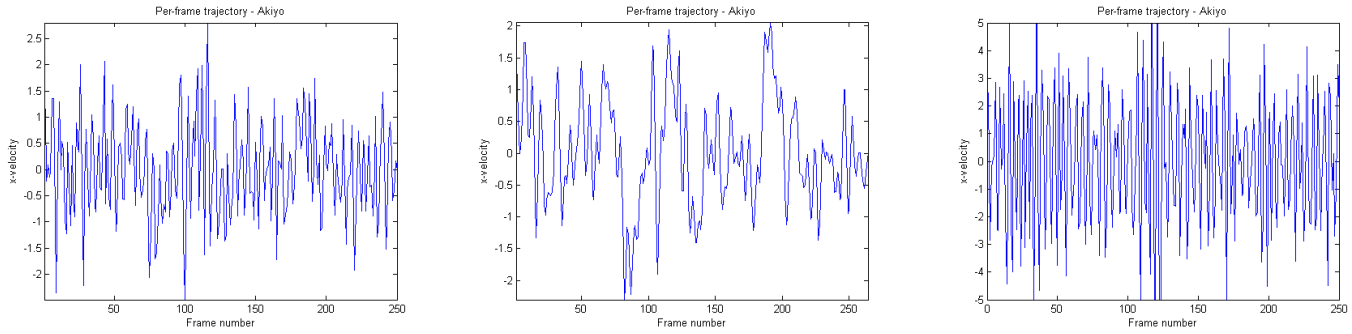
In the most basic scenario, this is due to the physiological tremor of the person holding the camera. Less noticeable causes of jitter include environmental factors such as ventilation outlets on the cinema screen, projector flickering and, if a stabilisation platform such as a tripod is used by the pirate, vibrations due to the activity of the other people in the room. In such cases, the high image resolution of modern acquisition devices provides an advantage to the forensic analyst, as smaller movements will result in displacement magnitudes beyond the subpixel range.

Referring to the diagram in Fig. 1, the recapture process adds jitter to the motion trajectories of image features over time. In the diagram, d_0 and d_1 are the distances between the camera optical centre and the opposite corners of the screen being recaptured. For minimal distortion, the screen is kept approximately parallel to the camera image plane, i.e. $d_0 \approx d_1$. In this case, the homogeneous 2D coordinates of the projection of any given 3D point $\mathbf{X}$ in the field of view at a distance d from the optical centre are:

$$\mathbf{P} \cdot \mathbf{X} = \begin{pmatrix} \frac{x}{d} & \frac{y}{d} & 1 \end{pmatrix}^T \tag{1}$$

Whenever $d_0 \approx d_1$ and the distance between the camera and the screen is much larger than the size of the screen, any jitter motion vectors added for each frame are approximately uniform throughout the frame due to uniform foreshortening. This is in contrast with noisy, jittery recordings of natural scenes, where the parallax effect will cause the motion of farther away portions of the scene to be perturbed to a lesser degree. Hence, while the jitter in itself indicates that the scene was captured by a handheld camera, it is the uniformity of the jitter throughout the frame that characterises the filming of a planar surface approximately parallel to the camera plane.

The problem of uniform jitter detection can then be decomposed according to the workflow shown in Fig. 2. First, given an input video, its visual features are tracked over time in order to create one 2D trajectory signal for each feature tracked. Since the added jitter is assumed to be a



(a) (b) (c)

Fig. 3: (a) Horizontal velocity signal from a tracked feature in the Akiyo sequence. (b) Its low-pass and (c) high-pass components.

relatively high frequency signal, each trajectory undergoes a 2-level wavelet decomposition after which only the high frequency components are retained. Finally, a normalised cross-correlation (NCC) matrix is populated by calculating the correlation between the high frequency components of all the trajectories. The motion and correlation histograms are then tested against known distributions trained from original and recaptured datasets by means of the Kullback-Leibler (KL) divergence.

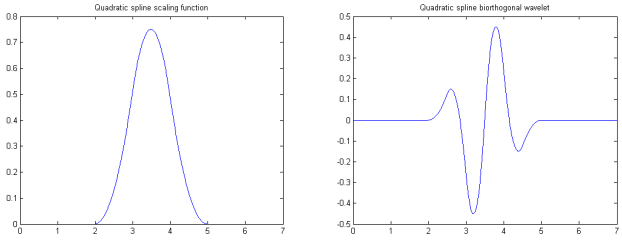
In the next subsections, each operation along the chain is examined in greater detail.

A. *Trajectory tracking*

The first phase of the algorithm aims to create 2D motion trajectories of the feature flow throughout the input sequence. To this end, features are first detected with the FAST feature detector [14] with subpixel refinement in the opening frame of the sequence. Typically, the number of features detected for tracking ranges between 500 and 2000 depending on the input image quality and resolution. The detection thresholds are manually adjusted so that features will be identified in approximately all areas of the frame, so that jitter uniformity can be tested for across the whole frame.

Detected features are then matched to their updated locations in the following video frame via a Pyramid Lucas-Kanade (LK) tracker [15]. The process is then iteratively repeated using the previously matched features as starting detected features to be matched in the following frame. Stored trajectories consist of a 2D motion vector per consecutive frame pair.

The choice of an LK tracker is due to its simplicity: a fast, effective tracker with no motion prediction phase makes sure that features are matched only when they are visible and their appearance remains unchanged. This guarantees a high matching confidence, and avoids the introduction of noise in our measurements from motion prediction or regularisers. Such high confidence comes at the price of matching rate, however for our application a few hundred successfully tracked features are normally sufficient. Trajectories with unsuccessful matches are removed from the final set.



(a) (b)

Fig. 4: (a) Quadratic spline scaling function and (b) its corresponding biorthogonal wavelet.

B. *Wavelet decomposition*

In order to separate the relatively high-frequency jitter components from the smooth local feature motion, each trajectory is input to a wavelet decomposition stage. First, the trajectories are converted from absolute coordinates to x-y velocities through a finite differences scheme. Then, a 2-level wavelet undecimated decomposition is performed using a quadratic spline and its corresponding biorthogonal wavelet as filters, shown in Fig. 4.

After the decomposition, the twice high-pass filtered trajectories' components are input to the next stage of the algorithm to quantify their uniformity. In Fig. 3 an example of a single feature velocity signal from the Akiyo video sequence is shown together with its low-pass and high-pass components.

C. *Correlation matrix population*

In this stage, the correlation between the high-pass components of the horizontal and vertical velocities is computed to test for their uniformity. Given N tracked features and their corresponding velocities v, an $N \times N$ normalised cross correlation matrix is populated where each element at location (i, j) is the result of:

$$NCC(i,j) = \sum_{i,j \in N} \frac{|v_i(t) - \bar{v}_i|\,|v_j(t) - \bar{v}_j|}{\|v_i\|\|v_j\|} \qquad (2)$$

The NCC measure is a good indicator of feature motion uniformity, and is therefore suitable for the proposed workflow. Given a natural scene captured by a static camera, feature velocity will be only locally correlated as isolated scene segments will be moving with common motion characteristics. Similarly, in most natural scenes only a fraction of the features is dynamic, with the remaining static zero-norm trajectories yielding very low correlation values. On the other hand, the jitter present in recaptured videos is by construction highly correlated as shown in Section III.

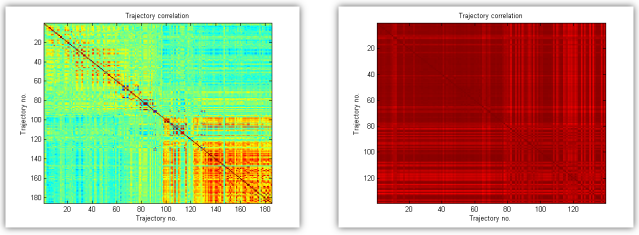
An example of this behaviour is shown in Fig. 5, where isolated patches of high correlation are found in the NCC matrix from the original video among a generally low correlation average. The NCC matrix from the recaptured video exhibits a uniformly high correlation values of its high-frequency velocity components.

D. Hypothesis testing

Qualitatively, using our previous considerations on jitter uniformity and local motion correlation it would be possible for a human observer to use the NCC matrices as telltale signs of recapture. However, in order to develop an automatic classifier the correlation distribution for both original and recaptured videos was considered.

For each of the 10 standard testing videos shown in Fig. 7 and their recaptured versions, a correlation distribution was created for their horizontal and vertical velocity components. The resulting 20 distributions per capture condition were then averaged together yielding two empirical distributions, representing the velocity correlation distribution of original and recaptured videos respectively.

From the graphical representation of the distributions in Fig. 6, the qualitative comments made earlier can be numerically observed directly. The correlation distribution for the recaptured case is mostly concentrated around the high-correlation end of the axis. For the non-recaptured case, there is a slight peak at the high-correlation end representing both local motion



(a) (b)

Fig. 5: Heat maps of NCC correlation matrices populated from the trajectories of (a) original 'Akiyo' and (b) recaptured sequences. Higher colour temperatures indicate a higher correlation value.

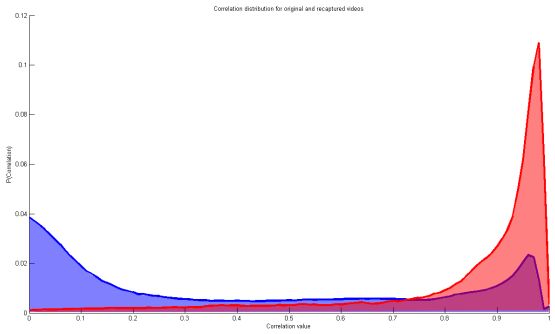


Fig. 6: Average distribution of NCC values from original (blue) and recaptured (red) videos. The distribution for recaptured videos is highly concentrated around the maximum value of 1, while original videos exhibit a much more uniform correlation, with maxima at either end due to local jitter and static features.

and noise from the tracker, while a more significant portion of the distribution is concentrated towards the low-correlation end of the graph.

Together with their differences in shape, the classifying power of the two distribution is also indicated by their non-overlapping area, which amount to 69.39% of their union. In the proposed workflow, these two distributions are previously created with labelled training video sequences. Incoming videos are classified by finding which one between the distribution for original videos P_{org} and recaptured videos P_{rec} minimises the KL divergence with the calculated input correlation distribution P:

$$\min\left(D_{KL}(P, P_{rec}), D_{KL}(P, P_{org})\right) \qquad (3)$$

where:

$$D_{KL}(P,Q) = \sum_{k \in [0,1]} P(k) \log \frac{P(k)}{Q(k)} \qquad (4)$$

In Eq. 4, $P(k)$ is the distribution of the correlation of the sequence to be classified, while $Q(k)$ is one of the two trained distributions. The output of the classifier is the category of the trained distribution that minimises the KL divergence with the input.

IV. Numerical Results

The proposed algorithm has been tested with 10 freely available test sequences[1], shown in Fig. 7. Each sequence was then recaptured with a compact Kodak V550 digital camera at a resolution of 640x480. The camera was held by hand as stably as possible to recapture the scenes directly from an LCD screen from a distance of approximately 1m, so that the picture would completely fill the field of view. Screen edges or other extraneous features were manually cropped in order to not provide static features that might aid the classification.

[1] All sequences available at: http://media.xiph.org/video/derf/

Fig. 7: Datasets used for evaluation. From left to right, top to bottom: 'Waterfall', 'Paris', 'Hall monitor', 'Container', 'City', 'Bridge', 'Akiyo', 'Salesman', 'Station', 'Old town crossing'.

From each sequence, a number of frames ranging from 100 to 500 was selected. The variable number of frames depends on the length of a sequence with stable features, i.e. without occlusions or features moving outside the field of view. For each of the 20 sequences, visual features were tracked with common set of tracker settings, yielding a number of trajectories ranging from 500 to 2000 depending on the image characteristics and video resolution.

The velocities obtained from the trajectories were then used to generate correlation distributions. The accuracy of the proposed method was then tested with a leave-one-out cross-validation. The results are reported in Table I.

	Classified as recaptured	Classified as original
Recaptured videos	90%	10%
Original videos	30%	70%

TABLE I: Classification accuracy results for original and recaptured sequences.

As the results suggest, the proposed method is effective for correctly identifying recaptured sequences. However, some errors are present in the classification of original sequences. Considering our dataset, such classification errors originate from the 'City', 'Station' and 'Waterfall' sequences. The first presents some significant challenges as it consists of an aerial view of a city recorded while looking down from a helicopter. The distance of the city from the helicopter is therefore approximately uniform for most features, while significant jitter is introduced during the original recording. The remaining two sequences both consist of a generally static scene being gradually zoomed out by a handheld camera. The apparent feature motion introduced during the zooming out process is highly correlated, and jitter is introduced by the operator.

V. Future directions

The testing conditions have also not been chosen in favour of the proposed technique: low original and recaptured video resolution results in coarser trajectory estimation from the tracker, unable to pick the finer differences in motion due to parallax. Moreover, feature-length videos do not consist of a single scene: multiple different sequences from longer videos can be sampled to check for consensus in the classification outcome.

Consequently, the first improvement to the proposed technique will be the automatic segmentation of the input video into multiple scenes. The scenes can then be analysed individually to look for jitter consistency (in terms of motion magnitude) across the different scenes, thus providing a recaptured video indicator that would not depend on prior training and possibly lowering the false positives rate.

Another extension to the method would consist of the use of a multi-scale tracker able to capture subtler motions. This, together with an adaptive decomposition of the trajectory signals based on their high-frequency content, might help the analyst in identifying sequences recaptured under stabler conditions, such as with cameras on tripods.

Finally, the improved method will be tested on more extensive amateurial datasets, such as YouTube videos, with particular attention to the false positive rate.

VI. Conclusions

In this paper a feature-based technique has been proposed for the task of automatic classification of recaptured videos. The method is based on high-level visual features and therefore robust to compression and common image processing operators.

The proposed technique has been tested with 10 freely available videos and their recaptured versions, with a classification accuracy of recaptured videos of 90%. Classification performance of original sequences can be improved by repeated subsequence sampling over the full video length.

Acknowledgments

This work was supported by the REWIND Project funded by the Future and Emerging Technologies (FET) programme

within the Seventh Framework Programme (FP7) of the European Commission, under FET-Open grant number: 268478.

REFERENCES

[1] Youtube statistics. [Online]. Available: http://www.youtube.com/t/press_statistics

[2] M. Goljan, J. Fridrich, and J. Lukas, "Camera identification from printed images," in *Proceedings SPIE on Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819, no. 1, 2008, p. 68190I.

[3] H. Yu, T.-T. Ng, and Q. Sun, "Recaptured photo detection using specularity distribution," in *Proceedings of the International Conference on Image Processing (ICIP) 2008*. IEEE, 2008, pp. 3140–3143.

[4] H. Cao and A. C. Kot, "Identification of recaptured photographs on lcd screens," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1790–1793.

[5] X. Gao, B. Qiu, J. Shen, T.-T. Ng, and Y. Q. Shi, "A smart phone image database for single image recapture detection," in *International Workshop on Digital Watermarking (IWDW)*, 2010, pp. 90–104.

[6] W. van Houten, Z. J. M. H. Geradts, K. Franke, and C. J. Veenman, "Verification of video source camera competition (camcom 2010)," in *Recognizing Patterns in Signals, Speech, Images and Videos - ICPR 2010 Contests*, 2010, pp. 22–28.

[7] W. van Houten and Z. J. M. H. Geradts, "Using sensor noise to identify low resolution compressed videos from youtube," in *Third International Workshop on Computational Forensics (IWCF)*, 2009, pp. 104–115.

[8] W. Wang, S. Member, and H. Farid, "Exposing digital forgeries in interlaced and de-interlaced video," *IEEE Transactions on Information Forensics and Security*, vol. 2007.

[9] M. Kobayashi, T. Okabe, and Y. Sato, "Detecting forgery from static-scene video based on inconsistency in noise level functions," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 883–892, 2010.

[10] W. Wang and H. Farid, "Detecting re-projected video," in *10th International Workshop on Information Hiding*, 2008, pp. 72–86.

[11] J.-W. Lee, M.-J. Lee, T.-W. Oh, S.-J. Ryu, and H.-K. Lee, "Screenshot identification using combing artifact from interlaced video," in *Proceedings of the 12th ACM workshop on Multimedia and security*, New York, NY, USA, 2010, pp. 49–54.

[12] M.-J. Lee, K.-S. Kim, and H.-K. Lee, "Digital cinema watermarking for estimating the position of the pirate," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 605–621, 2010.

[13] M.-J. Lee, K.-S. Kim, H.-Y. Lee, T.-W. Oh, Y.-H. Suh, and H.-K. Lee, "Robust watermark detection against d-a/a-d conversion for digital cinema using local auto-correlation function," in *Proceedings of the International Conference on Image Processing (ICIP)*, 2008, pp. 425–428.

[14] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision (ECCV)*, vol. 1, May 2006, pp. 430–443.

[15] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," Intel Corporation - Microprocessor Research Labs, Tech. Rep., 2000.