On the impurity of street-scene video footage

Henderson, C; Blasi, S; Sobhani, F; IZQUIERDO, E; International Conference on Imaging for Crime Detection and Prevention

On the impurity of street-scene video footage

Craig Henderson, Saverio G. Blasi, Faranak Sobhani and Ebroul Izquierdo

Multimedia and Vision Lab, Queen Mary University of London
{c.d.m.henderson, s.blasi, f.sobhani, ebroul.izquierdo}@qmul.ac.uk

Keywords: Video Corruption, Image Quality, File Formats

Abstract

In this paper, we present the technical challenges facing researchers in developing computer vision techniques to process street-scene videos *from the wild*. Video footage captured by surveillance CCTV cameras and hand-held devices such as mobile phones and body-mounted cameras worn by police officers pose particular difficulties. Video formats are varied and often non-Standards compliant which leads to apparent corruption when rendered using standard players. Footage is low-quality either in resolution or in sharpness caused by free movement of the camera, fast panning and zooming or weather conditions. We describe our experiences working with the Metropolitan Police in London to find a solution to these problems and enable computer vision techniques to be used in the forensic analysis of videos in criminal investigations.

1 Introduction

The Metropolitan Police in London (the *Met Police*) have found that the opportunity to use computer vision technology in the analysis of real-world street-scene video is severely limited because of the practical constraints in the variety and poor quality of videos available to them. Consequently, in a large criminal investigation, police forces employ numerous officers and volunteers to watch many hours of camera footage to locate, identify and trace the movements of suspects, victims, witnesses, luggage and other inanimate objects. Their goal is to piece together a story of events leading up to an incident, and to determine what happened afterwards. For example, a recent large-scale missing person investigation by the Met Police obtained 8 days (8×24 hours) of continuous video camera footage from local authority street cameras which amounted to 30 Terabytes of video data. In addition to this was footage for shorter durations obtained from private residences, shops and other businesses. Such a large amount of video is difficult to manage, and studying long periods of footage is time consuming and intense work for the officers and volunteers involved.

In this paper, we highlight the practical challenges faced by a police investigation team in analysing these videos. We report our experience in processing three terabytes of video supplied to us by the Met Police as a part of the European LASIE project[1] that aims to *significantly increase the efficiency of current investigation practices by providing an automated initial analysis of the vast amounts of heterogeneous forensic data that analysts have to cope with.*

[1] http://www.lasie-project.eu/

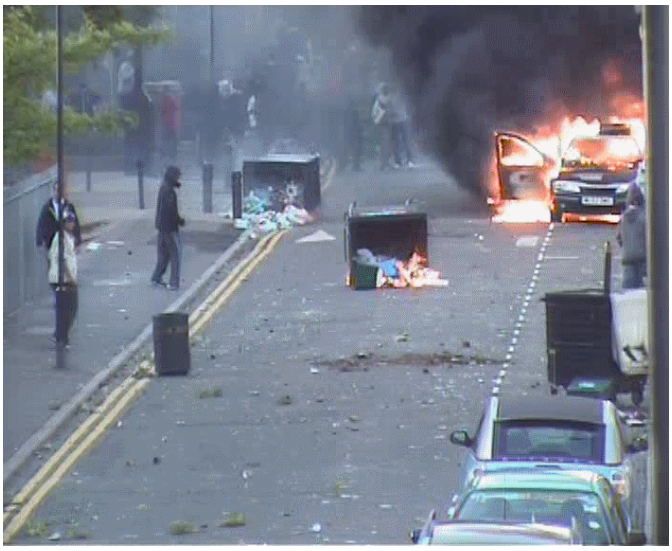


Figure 1: Civil riots in London, UK during August 2011 were captured on many hours of video footage, featuring many criminal offences in progress. Manual analysis of all the video is very time consuming, but the video quality is challenging for current computer vision algorithms to process with robust results.

2 Context

Empirical research algorithms are typically demonstrated to work with high quality video. Two common examples are Hollywood movies *Groundhog Day* (Ramis, 1993) and *Run Lola Run* (Tykwer, 1998) used in [1, 15, 16] and subsequent comparative papers, or *Casablanca* (Curtiz, 1942) in [16]. These high quality videos have a high frame rate and good image resolution. The context in which street-scene videos are recorded differ in a number of significant ways from a feature film and produce scenes that are challenging to computer vision algorithms.

Long-running video sequences The innovation of [15] to apply text retrieval theory and practices to video searching defined *visual words* for describing structure in images. The method tracks features across *shots*, a contiguous sequence of

frames taken from a single camera within a scene. The number of frames and discovered features to process is manageable because of the relatively short period of time covered by a shot. The average shot length in feature films was 8-11 seconds before 1960 and had reduced to 4-6 seconds by 2006 [3]. Localised processing of feature movement is an aid to the algorithm by reducing the data volumes and providing a natural delineation of processing. Where necessary, cross-shot feature tracking can be considered at a later processing stage once features have been tracked within a shot. Boundary shot detection is well documented as an important prerequisite step to automatic video content analysis [19] as shots are regarded as the basic unit to organize the sequenced content of video and primitives [2].

Videos used in criminal investigations are very different. Fixed surveillance cameras fall into three categories; those that do not move and continuously record the same field of view, automated movement cameras that follow a defined motion such as a figure-of-eight to try to maximise area coverage, and human operated cameras that can be pivoted up and down, rotated around 360° and zoomed to varying depths[2]. Each of these cameras produce a video consisting of a single shot that can last for hours. Body-mounted cameras and mobile phone footage also produce uninterrupted video sequences that can last several minutes, and hundreds of frames. Without a natural delineation of shot change, contemporary methods of object tracking and mining become less manageable, demanding large computing resource to process.

Camera movement In static surveillance cameras the focal length and field of view are both fixed, and do not follow any activity. A car or a person that subsequently becomes of interest to police does not stay within shot, or even within focus. These fleeting glances can be important to an investigation but could easily be missed by reviewers scanning many hours of CCTV video.

An alternative to static cameras are those which passively record following a pre-defined motion path, with the camera mounted on a bracket that automatically moves around a loop or figure-of-eight to maximise the coverage of an area with a single camera. Objects will move in and out of view regularly within a sequence of frames. Other cameras are human-operated and can record very erratic movement with dramatic changes of focus and rapid zoom as the camera operator wrestles with the controls to record action on the streets. Individual frames can therefore be very blurred. The fast movement in pan and zoom, either in the manually controlled camera or to a lesser extent in a fixed-path motion camera degrades the image quality further, and is somewhat unique to the security videos such as those that we analyse.

Environmental Security cameras record in uncontrolled environments. The footage is continuous, without any controlled change in focus, lighting and position. As a result, images have poor colour clarity and little discriminative or representative texture definition.

Many variations occur over a long-running video sequence. The sun changes through the day in position and intensity, and at night the scene changes to artificial ambient lighting and spot lighting from vehicle headlights, for example. The quality of images from each security camera therefore varies considerably, and this inconsistency can cause difficulties in finding correspondences in images even from the same camera.

Variations in weather over time cause very different images to be captured by a camera at different times. A change from sun to cloud affects the light intensity and colour definitions within the image. Rain or snow can appear as noise and even occlusions in extreme conditions. Fluctuating lighting conditions can also be caused by burning fire and by emergency vehicle lights, especially at night, and are commonplace in video that undergoes forensic analysis.

Closed-circuit television (CCTV) cameras are often sited very high and cover a long field of view where objects in the distance lack colour definition and texture clarity and can be difficult to identify even for a human. Fast camera movement pan or zoom, frenzied motion within a frame, or a combination of both can cause significant blurring in frame images which results in a lack of texture. Camera instability in free-hand or body-mounted cameras cause serious image blur and erratic movement.

Video acquisition and recapturing The source of video footage used in a police investigation can be varied, as there is a lack of standardisation in CCTV systems. Obtained footage is often in a proprietary format that can only be viewed on-screen by a manufacturer-supplied application, and the flexibility and usability of these applications vary tremendously. To achieve their goal of forensic analysis and examination of segments of video, and to be able to edit videos into a *story* that can be used in a criminal court, the Met Police have employed creative solutions to overcome the limitations of the source video images. The result is a tedious and time- and resource-intensive activity to transcode the video footage by re-capturing the video as it is played on a computer screen. The resulting video file is in a standardised format that can be viewed and edited as required, and can also then be used in computer vision applications and research.

A consequence of the difficulty of acquisition is that the standardised video is often without meta-data which may have been useful, such as the video frame rate and date/time stamps. These difficulties contrast with environments in most research which use Hollywood films with fast, and known, frame rates, high resolution images with consistent lighting, and where scenes are repeatedly re-shot until the quality meets an acceptable standard.

A further complication with the frame rate is introduced by the recapturing process. Re-capturing records at a fixed frame rate, perhaps 25 fps. If the video being played is at a lower frame rate, then multiple frames will be captured for each frame in the original video file. The playback is visually unaffected, but this duplication of consecutive frames adds another complication for computer vision applications as the

[2] these cameras are called *PTZ*, reflecting their capability to *Pan*, *Tilt* and *Zoom*

amount of movement between pairs of adjacent frames is inconsistent. A second piece of meta-data is time sequence data. Time sequences would enable software to be able to synchronise video captured from multiple cameras, for example, based upon the time information associated with the video sequence. Edelman [9] reported on a system at the Netherlands Forensic Institute which uses Optical Character Recognition to read video timestamps from the video frame images. Such a technique is not reliable enough to provide sufficient meta-data for steering Computer Vision algorithms, however; the Met Police observe that camera timestamps are unreliable as the accuracy of the time is dependent on the ongoing maintenance of the CCTV system, and varies considerably between local authority, police and private owners of surveillance systems. Standard police procedure now is to record the actual current time and the presented CCTV time when a security video is seized for an investigation. This enables the police to calculate the offset of the CCTV time, but is fragile to the system clock having been altered since footage of interest was recorded.

3 Visual image quality

In contrast to Hollywood movies, CCTV cameras videos vary considerably in their frame rate and image resolution. Established methods of feature detection, extraction and matching perform less well on these videos than on high-definition images with sharp focus and controlled lighting conditions [12].

The frame rate of a video is measured by the number of frames per second, *fps*, that are recorded. With a lower frame rate, the time between frames is greater, features are further away relative to the previous frame and move greater distances relative to each other. Adjacent frames therefore have a greater visual difference than those from a high frame rate video. This difference can significantly affect the robustness of computer vision algorithms that often rely on the *a priori* knowledge that two adjacent frames in an video are very similar. As an example, a feature tracking algorithm makes the determination of whether features are related or not based on the amount of global and relative movement between frames, known as *spatial consistency*. In a low frame rate video, such determination becomes less robust as the movement threshold must be increased to compensate for the additional movement, and this can introduce noise and mis-classifications. It would be possible to configure spatial consistency algorithms using a video's meta-data, for example to adapt the spatial distance threshold of related features based on the frame rate of the video. In our area of interest, surveillance videos very often have no associated meta-data, and cannot therefore be used as a reliable input into algorithmic choices for spatial consistency parameters.

CCTV cameras vary considerably in their frame rate and image resolution. Low frame rates reduce the number of images that make up the video sequence and a low resolution reduces the size of each video frame. Together these two attributes can significantly reduce the amount of storage required, and therefore the cost of storing the captured video and so are often reduced by organisations who seek to minimise the overhead of their security operations. The clarity of images from different security cameras also vary considerably, and this inconsistency can cause difficulties. Images are often low resolution with poor colour definition and have little discriminative or representative texture definition, and images from these need to be matched with those from higher definition images. Quality is further reduced by varying weather conditions where the changes in light, presence of rain, snow, mist or fog, direct sunlight and shadows can all affect the image, and the ability for a feature extractor to consistently describe an image region.

4 Where contemporary algorithms fail

The position of a region in one frame with respect to within an adjacent frame is described by a simplified linear model [18]

$$\begin{bmatrix} m_x \\ m_y \end{bmatrix} = \begin{bmatrix} s & -\theta \\ \theta & s \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix} + \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \tag{1}$$

where (m_x, m_y) is a motion vector at position (x, y), (x_g, y_g) is the centre of gravity of the region, and $(\theta, s, \delta x, \delta y)$ are motion parameters that can be estimated using the steepest gradient decent algorithm [13]; θ is rotation, s is scale, and $(\delta x, \delta y)$ are translation parameters. The state of an object at time t can be described by a vector that consists of six parameters of an affine transformation [17]

$$\mathbf{x}_t = (\delta x_t, \delta x_t, \theta_t, s_t, \alpha_t, \phi_t) \tag{2}$$

where $\delta x_t, \delta y_t, \theta_t, s_t$ are as above, and α_t and ϕ_t denote aspect ratio and skew direction, respectively. If an object is at a known position (x, y) in frame $t-1$, then its position in frame t can be predicted by [17]

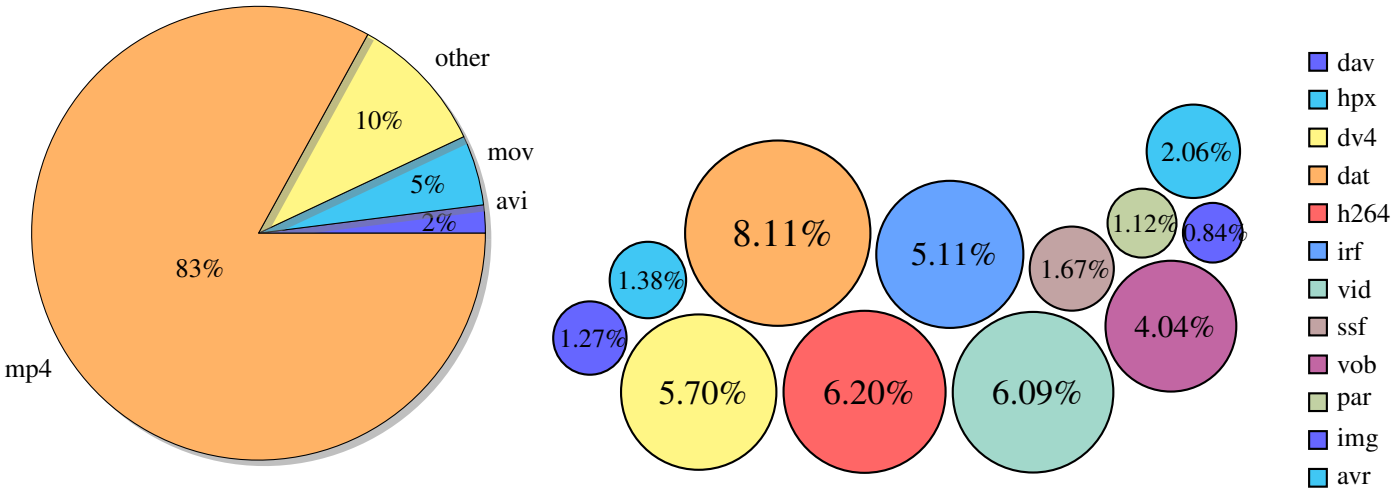
$$\mathbf{x}_t = \mathbf{H}_n \mathbf{H}_p \mathbf{x}_{t-1} \tag{3}$$

where $\mathbf{H}_n$ is a matrix calculated from Gaussian noise and $\mathbf{H}_p$ is a prediction matrix. If $\mathbf{H}_p$ is the identity matrix, then the previous frame $\mathbf{x}_{t-1}$ is used as the prediction for the current frame, as is the case in much of the literature.

Motion estimation techniques such as *Mean Shift* [5, 10] and CAMshift (*Continuously Adaptive Mean Shift*) [4] use *a priori* knowledge that the object being tracking will move only a short distance between frames to reduce the search area and search for objects within small spatial variation limits [14]. They have therefore been used successfully in real-time tracking applications [7, 6], but are less effective in CCTV video sequences with low frame rates because the movement between frames is non-deterministic and too large.

5 Video formats

Among CCTV security manufacturers, there is no industry standard for resolution, frame rate or file formats. AVI is a commonly used container for storing videos, but most of the CCTV manufacturers are not compliant with encoding standards of the video data stream contained within. Hardware and software specifications used in digital video recorders (*DVR*s) – file formats, encoding and compression – are not strictly adhered to, causing observable *corruption* when the video is



(a) The three most frequent file types are MP4, AVI and MOV, which account for 90% of the files in our reduced dataset

(b) Breakdown of the remaining 10%

Figure 2: Breakdown of *readable* file formats in 3.07 Tb of data from criminal investigations. Readable data accounts for only 929 Gb of the total available. The percentage of files of each type in our dataset is shown.

played using standard players. A variety of informally documented and low-cost DVR systems exist in the current market with unreliable hardware and poor Standards support.

A digital video recorder is a stand-alone unit with the capability of storing sequences of images on a hard disk drive. There are countless DVR manufacturers, each producing many products with a variety of features. As for security, a true DVR is recognised as a sophisticated system combined with specifically designed hardware, software and sub-assemblies containing built-in checks and balances in order to interact with each other to create a robust solution. A DVR is central to a CCTV surveillance system and its quality dictates the quality of the system as a whole [11]. A high-quality CCTV system in a well-lit environment connected to a good DVR will record a good quality image, but the same system connected to a low-cost DVR with insufficient processing power, faulty hardware encoder or poor codec algorithms will record inferior images.

Of the 3.07Tb of video data provided to us, only 929Gb is in a non-proprietary format. This subset of files consists of a number of file formats (Figure 2). However, although the files purport to be Standard formats, viewing many of the videos using standard video players demonstrate incorrect aspect ratio, half screen corruption, long sequences of blackness, missing key frame or upside-down videos (Figure 3). This is because most CCTV manufacturers tend to use proprietary video codec schemes [8] instead of relying on widespread video coding standards.

6 The case for proprietary CCTV encoding

Compression methods are typically optimised to encode multimedia and broadcasting signals, whereas CCTV footage has very characteristics and requirements.

Costs Broadcasting companies usually make use of high specification hardware solutions to encode signals at very high qualities. In the case of CCTV footage, the encoding needs to be performed on digital signal processing (DSP) boards with very limited resources.

Real-time transmission Most multimedia video coding standards require processing of a large group of consecutive frames before encoding, which introduces relatively long delays. These delays are tolerated in broadcasting applications (even in *real-time* TV transmission, where the signal is always subject to a delay in the order of seconds). On the other hand, CCTV footage needs to be transmitted in real-time with the smallest possible delays, to allow a prompt response.

Scalability is often a desired component of CCTV schemes. Ideally, a CCTV coding scheme should offer the possibility of having multiple bitstreams at different bitrates (and correspondingly different qualities), to allow the receiving end to select quality depending on bandwidth availability. While scalable video coding solutions exist in the market, they are typically not tailored to CCTV footage and require very complex encoding schemes to achieve the desired level of scalability.

Embedded elements Video coding standards usually treat any other multimedia information (i.e. audio, text, metadata) as signals which need to be encoded and transmitted separately. In the case of CCTV schemes, however, it is often desired to embed this information within the bitstreams. This prevents de-synchronisation issues and allow for fast processing of the received data. Moreover, many CCTV codecs embed a timestamp within the signal: this is crucial to recover the exact time and day when the signal was captured. See §2 for challenges in relying on this for computer vision applications.

Manufacturer	Player	File Extensions	Company URL
Avtech Software Inc.	PlayerLiterHJ	dv4, .vse .vs4, .avc	http://www.avtech.com.tw/
Cop Security	COPPlayer	.arv, .har	https://www.cop-eu.com/
Dedicated Micros	NetVu VCR	.par	https://www.dedicatedmicros.com/
IDIS	Embedded Clip Player	.exe	http://www.idisglobal.com/
GeoVision	GVSingle	.avi	http://www.geovision.com.tw/
JLE CCTV	iFile Playpack	.irf	http://jlecctv.com/
Speco Technologies	DAV file	.dav	http://www.specotech.com/
Synectics System Group Ltd	FSM Player	.dat	http://www.synecticsplc.com/
Sensormatic	Intellex Player	.img, .im	http://www.sensormatic.com/
Vista	Smart Player	.hta, .hi, .hpx	http://vista-cctv.com/
Vista	Quantum Plus H264 Player	idx, vid	http://vista-cctv.com/

Table 1: Proprietary video players used in our sample dataset

Multiview and multiple cameras It is common for video surveillance systems to include multiple cameras or multiview architectures. CCTV codecs often compact these signals into a single bitstream, to reduce the storage needs required to compress the sequences. Proprietary decoders also allow for direct switching among different cameras while displaying.

While proprietary standards ensure that signals can be compressed and transmitted according to the requirements of a video surveillance scheme, they also represent a serious issue when sharing and displaying the signals, as illustrated in the rest of this paper.

7 Video corruption

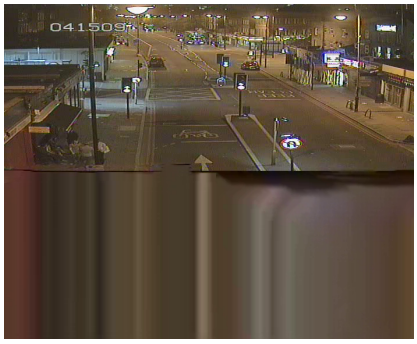
With the term *video corruption* we refer to all problems which limit the viewer from displaying the sequences, or more generally have an impact on the quality of the content. As such, we will also present some examples in which the original files are not *corrupted* in the technical meaning of the term, but in which the nature or format of the files can limit their accessibility and consumption. There are many causes of corruption in CCTV footage and in fact video corruption is an extremely widespread problem when dealing with this kind of content. The reasons for this may reside in the typically adverse conditions of capturing discussed earlier, or they may simply be due to incorrect encoding/decoding of the video files which may affect the displaying in unpredictable ways. The latter problem is especially relevant because most CCTV manufacturers tend to use proprietary file formats/video codecs.

While an exhaustive classification of the corruption issues we encountered in the available content is very difficult, we propose to categorise these into three classes, depending on the stage during which the issues affected the file. In particular we refer to *source corruption* issues where the content got corrupted at the source, before being compressed, encoded and transmitted. We refer to *coding corruption* issues where the content got corrupted during the encoding stage which means it cannot be decoded, or it is wrongly decoded. Finally we refer to *format corruption* issues where the content is represented in an atypical, proprietary format or compression standard which is difficult to manipulate or can only be displayed using proprietary software.

Source corruption can arise for many reasons, due to issues with the capturing device such as distortion of the lenses or problems with the digital sensor, and as such it is irreversible. An example of such corruption problems can be seen in Figure 3a. One frame of a video sequence is displayed in the figure. Clearly the video shows some corruption affecting the aspect ratio of the sequence (704×288). Initially we thought the problem might be due to decoding issues, or wrong displaying settings. Unfortunately a more thorough analysis of the video file revealed that the file was already distorted before being originally encoded (using the AVC standard). This means that it is physically impossible to recover the correct aspect ratio of the video, unless manually inspecting the file. We refer to *source corruption* in such cases when the lighting conditions at the time of capturing are such that little or no portions of the sequence are visible. This is a surprisingly common problem arising with videos taken during night time. There are tools in the market which aid recovering some information, for example the *MPEG4 Modifier* tool enables a change of aspect ratio or luminance without re-encoding. Unfortunately most of these software only work on specific formats – for instance, MPEG4 Modifier only works for videos encoded in the MPEG4 format. Moreover, the typically very low resolutions of the captured sequence make the use of these tools very challenging, with mixed output results.

Coding corruption can arise either because of problems during the compression of the signals, or because the bitstreams (i.e. the compressed signals) got corrupted while stored, copied or transmitted. An example of such corruption can be seen in Figure 3b. The figure shows a frame decoded from a sequence encoded using the AVC standard in its main profile. When decoding the sequence, conventional AVC decoders complain due to some missing syntax elements in the streams. Thanks to the relatively high flexibility and two figures complex error

(a) Incorrect aspect ratio



(b) Half-screen corruption

(c) Almost black video

(d) Missing key frames

(e) Upside-down video

Figure 3: Examples of video corruption common in security videos from uncontrolled sources

resilience mechanisms of AVC decoding, such problems still allow for the file to be decoded and displayed, but the decoded signal is clearly not displayed optimally. Again, these tools only work on specific formats or standards.

Format corruption problems arise where the sequences were compressed and encoded using proprietary codecs, as illustrated in the previous section. Police forces are often presented with a variety of encoded bitstreams, each requiring different decoders and players. This creates several issues. First, the signals can be displayed only where the corresponding decoder or player can be installed on the local machine. Second, any processing requires a transcoding from such proprietary standard to existing universal standards, which may or may not be possible using software provided by the codec developers. Where this is not supported, the signals are often recaptured during display and converted to a new format (§2). This process is incredibly slow and computationally complex, and introduces high amount of noise in the produced signals. In the case where the files get corrupted or unreadable, it is impossible to recover information using any of the existing recovery or error resiliency tools available for universally adopted standards.

8 Conclusion

Our work with real-world security videos acquired from the original sources has shown that such video sequences present numerous difficulties for automated processing. These difficulties are often overlooked or not acknowledged in research literature. The presumption that reasonable quality videos are available for analysis is invalid for practical applications, and we have described a number of causes and potential resolutions from our own experiences.

Acknowledgements

This work is funded by the European Union's Seventh Framework Programme, specific topic "framework and tools for (semi-) automated exploitation of massive amounts of digital data for forensic purposes", under grant agreement number 607480 (LASIE IP project). The authors also extend their thanks to the Metropolitan Police at Scotland Yard, London, UK, for the supply of and permission to use CCTV images.

References

[1] Arasanathan Anjulan and Nishan Canagarajah. A unified framework for object retrieval and mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):63–76, January 2009. 2

[2] MN Muhammad Nabeel Asghar, Fiaz Hussain, and Rob Manton. Video Indexing: A Survey. *International Journal of Computer and Information Technology*, 3(1):148–169, 2014. 2

[3] David Bordwell. *The Way Hollywood Tells It: Story and Style in Modern Movies*. University of California Press, 2006. 2

[4] Gary R. Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No.98EX201)*, pages 214–219. IEEE Computer Society Press, 1998. 4

[5] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995. 4

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003. 4

[7] Dorin Comaniciu and Peter Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 2002. 4

[8] Matthias Doring. Requirements for the ideal CCTV video codec. Technical report, Geutebruck Technical Report, 2006. 5

[9] Gerda Edelman and Jurrien Bijhold. Tracking people and cars using 3D modeling and CCTV. *Forensic science international*, 202(1-3):26–35, October 2010. 2

[10] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, January 1975. 4

[11] K Graželis, D Barkauskas, and P Laptev. Minimum Requirements for Videos Surveillance System for Public Areas. 5

[12] Craig Henderson and Ebroul Izquierdo. Robust Feature Matching in the Wild. In *Science and Information Conference*, London, 2015. IEEE. 3

[13] H Nicolas and C Labit. Motion and Illumination Variation Estimation Using a Hierarchy of Models: Application to Image Sequence Coding. *Journal of Visual Communication and Image Representation*, 6(4):303–316, December 1995. 4

[14] Omar Oreifej, Ramin Mehran, and Mubarak Shah. Human identity recognition in aerial images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 709–716. IEEE, June 2010. 4

[15] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003. 2, 2

[16] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):591–606, 2009. 2

[17] Li Sun and Guizhong Liu. Visual Object Tracking Based on Combination of Local Description and Global Representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):408–420, April 2011. 4, 4

[18] Demin Wang. Unsupervised video segmentation based on watersheds and temporal tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):539–546, 1998. 4

[19] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A Formal Study of Shot Boundary Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17, 2007. 2