A pipeline for analysing image and video material in a forensic context with intelligent systems

Svenja Preuß[1] and Dirk Labudde[2]

Abstract: Shows like CSI seem to convey a certain view on the capabilities of forensic science based on the vast progress of digitalisation and the new technology, that goes along with it. But those depictions can be misleading and hardly represent a realistic reflection of reality. Nevertheless, this representation influences the public view on digital forensic analysis. This phenomenon is also known as the CSI effect. To present a more realistic view on practices in digital forensics, we want to introduce typical image and video analysis methods used in tackling real life forensic challenges and point to their capabilities as well as their limits. In this context an important area is image and video enhancement. With methods such as Super Resolution, images can be scaled up, their corresponding resolution gets enhanced and image noise can be reduced. During the subsequent image analysis, methods ranging from purely cognitive analysis to extraction of raw pixel values and various semantic information from the image or video, utilizing AI frameworks, are used. This allows for example to detect and analyse faces up to whole people, as well as objects in images or videos.

Keywords: forensic science; computer vision; AI frameworks

1 Introduction (CSI effect)

The profession of forensic scientist is gaining increasing interest among the general public [LS17]. Not least due to the numerous movies and shows, such as "CSI: Crime Scene Investigation" or its spin-offs "CSI: Miami", "CSI: NY" and other related shows. In these, scientific analyses and analysis techniques are increasingly moving in focus [LS17]. However, these analyses often hardly represent a realistic reflection of reality and are usually exaggerated, presented in a non-scientific style. Especially in the analysis of images and videos, sections can be enlarged repeatedly until even the smallest trace and detail is completely sharp visible. This results in a phenomenon which is heavily discussed in literature, the "CSI effect". [LS17] This phenomenon manifests itself in three different ways [PO05]. One, as alluded to earlier, is the increasing public interest and awareness of forensic science [PO05]. This is also reflected in the increasing demand for university programmes or vocational trainings in forensic science [LS17].

The attributed influence of the "CSI effect" on judges as well as other legislatives is

[1] University of Applied Sciences Mittweida, Department of Natural and Computer Sciences, Technikumplatz 17, 09648 Mittweida, Germany preuss2@hs-mittweida.de
[2] University of Applied Sciences Mittweida, Department of Natural and Computer Sciences, Technikumplatz 17, 09648 Mittweida, Germany labudde@hs-mittweida.de

doi:10.18420/inf2022_09

clearly more critical. The portrayal of forensic science in these shows creates an exaggerated expectation of courts for forensic analysis. Also, this expectation is reinforced by the displayed infallibility of the results of forensic analyses, creating a fear that all results will be accepted without question. [PO05]

Literature disagrees if the "CSI effect" has an actual impact on the legislative system or it is just a spurious correlation exaggerated by current findings. But it is undoubted that the shows have some influence on the point of view of forensic science in society, its methods and methodologies [LS17]. In order to draw a more realistic picture of digital forensics, especially with regards to digital multimedia forensics, some methods, which are applied in this field, will be presented here and evaluated on case studies.

2 Definition of forensic image processing and image analysis

In this section, the necessary terms for forensic image and video analysis shall be introduced and delimited from each other. Therefore, orienting on the work of Grimm F. [GR90], a distinction is made between **image analysis** and **image processing**. It should be noted that a video is nothing more than a sequence of images that are sequential in time, called frames [NI20]. Therefore, before any processing and analysis steps, the videos are separated into their individual frames. During image processing the original image is transformed so its information layout is more suitable for the respective analysis. A part of the image processing is the **image pre-processing**, whereby most of all the quality of the image should be improved. **Image segmentation**, likewise a subfield of the image processing, deals with the separation of the image into different image areas or its components. [GR90] In this way, relevant areas can be extracted and viewed independently of their surroundings. Image analysis, on the other hand, does not transform the image, but describes the content and semantic of the image in a certain way [GR90]. In the following, the methods of the individual field, which are used in daily work of a digital forensic scientist, will be displayed.

3 Image processing

3.1 Image pre-processing

One of the most frequent requests, which often appears in inquiries, is to improve the image or video quality. Often the material to be analysed is from security cameras located randomly near the crime scene, with the area of interest occupying only a small part of the image. Here we return to the example from the introduction, where a section of the image can be enlarged infinitively. An unlimited enhancement of image segments is not possible, since the resolution of an image also determines the information that can be displayed. Super resolution methods can be used in this context, to generate a higher-

resolution images from low-resolution ones, with the same content. However, it should be emphasized here that no new information can be gained, but only fill values can be estimated by linking existing information (interpolation). Two day to day used types of super resolution are on one hand basic interpolation and on the other hand methods that can be summarized as digital neural image enhancement technologies (DNET). The latter technologies not only increase the resolution, but also include other methods of image pre-processing, such as noise reduction or contrast enhancement.

Interpolation attempts to approximate unknown points of a function that lie between two or more known function values (samples) [FA14]. In terms of an image, the samples would each represent one pixel of the original image and the values to be determined would be the pixel values between these given pixels needed for magnification. Interpolation methods for calculating the intermediate values, which are mainly used here, are linear and cubic or bi-linear and bi-cubic. The difference between the former and the latter is that using the first two methods the interpolation is calculated in one dimension only and using the latter over two dimensions. Consequently, in bi-linear/bi-cubic interpolation, the nearest pixels are considered in both x and y directions. Usually, this two-dimensional interpolation is implemented by first interpolating in the x-direction and then in the y-direction [FA14]. As the name suggests, (bi-)linear interpolation is based on a linear function. In bi-linear interpolation the weighted average of the four nearest pixels is taken into account [FA14], all other pixels have no relevance. (Bi-)cubic interpolation is based on a third-degree polynomial function and considers the weighted average of the 16 nearest points [FA14], with closer points having a higher weighting [PMG13]. Cubic interpolation usually achieves better results than linear [FA14] [PMG13], but requires slightly more computation time [PMG13].

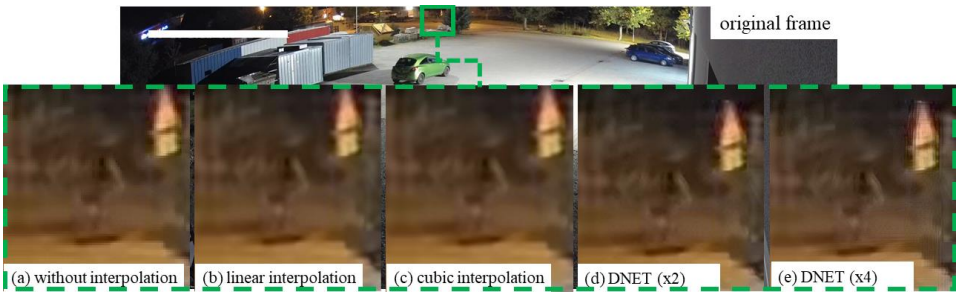

Fig. 1 Super resolution enlarged sections (a - e) from one frame (original frame) of the video footage from a security camera located near the scene of an attempted arson. The image sections show a person with a dog which had to be analysed.

Digital neural image enhancement technologies, on the other hand, are based on machine learning. They increase the resolution by learning the weights, based on data sets, to determine the information to be generated between the surrounding points. Consequently, this method generates the most probable weights of the surrounding pixel values based on the learned data. Currently, "neural-enhance", a free software for

enhancing images by Cardinale F. is applied [CA22]. In contrast to conventional Interpolation, these techniques have a probability of constructing information that was not originally present. As an example, a neural network trained on faces might generate a birthmark on the face even though there is actually no birthmark at that location, based on the trained data it is only very likely that there might be a birthmark.

Stacking is another way of enhancing visual perception, if several frames or images of the same scene or object are available. In the following, a distinction is made between **focus stacking** and **temporal stacking**. In focus stacking, individual images with differently sharp zones are combined to form a new image with a greater depth of field [KN19]. Temporal Stacking can also reduce noise [AT13] and merge overlapping image information by superimposing the individual frames. Temporal stacking is performed by several steps proposed by K. Atanassov [AT13]. Basically, a reference image has to be selected first, tending to choose the image with the highest information content and the lowest image noise [AT13]. Afterwards, key points are determined on the object to be stacked, which can be detected unambiguously and on the basis of which the frames can be aligned with each other [AT13]. Finally, the frames are combined with each other. Usually this is done by calculating the mean or average or by finding the median, for each pixel value across all frames. This can reduce random noise by producing constant signals more clearly.

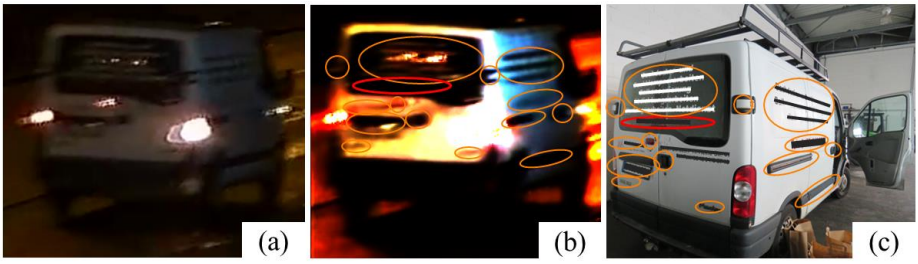

Fig. 2 (a): Reference image of a transporter captured by surveillance cameras at a local gas station. (b): Image stacked from six brightened frames of the transporter aligned using prominent keypoints of the transporter. (c): Image of the comparison transporter. The red and orange circles symbolize the individual feature for the correspondence analysis

However, stacking is only applicable if sufficient variable material is available. Furthermore, no non-existing information can be extracted here either. If material is only available in very poor resolution or quality, often insufficient information can be extracted. If temporal stacking does not provide adequate information, objects from different images whose similarity is to be investigated can also be compared by aligning them and fitting them to each other. Based on this, a later correspondence analysis (see section 3. step: "purely cognitive analyses") can be simplified. This derived form of stacking can be called **information merging**.

Stacking is already used in many different areas, for example biology, especially included microscopy [KN19]. Nevertheless, these methods can also be applied in a

forensic context.

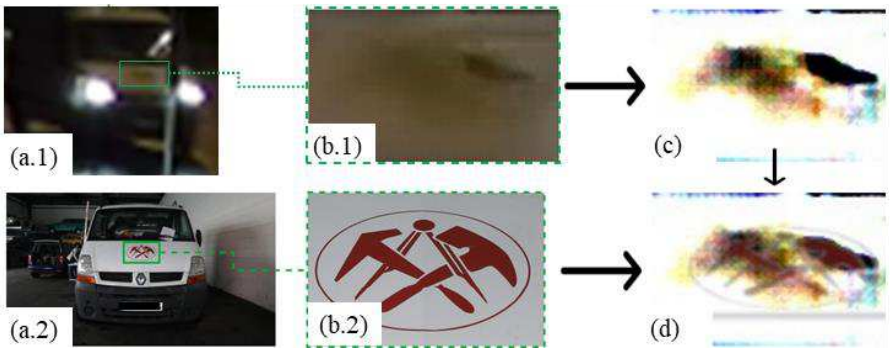

Fig. 3 Pipeline for information merging. (a): Transporters to be checked for matching. (b): Respective extracted area on the hood of the transporters (a.1) and (a.2). (c): Separately contrast and brightness adjusted extracted area on the bonnet of the transporter of (a.1). (d): Result of information merging.

Other improvements that can be used individually for the image and video material are the **adjustment of contrast** and **brightness** as well as **saturation**. Often a **white balance** is also necessary [BSS17]. **Colour normalization** is the adjustment of the colour histogram, which is a representation of various statistical information about the distribution of colour and brightness values [BSS17] in an image. By normalizing this histogram, colours can be balanced [IQ10] and a possible cognitively perceptible colour shift can be corrected. In addition to the above-mentioned stacking, **noise**, as described in [BSS17], can be **reduced** by smoothing the image, but always at the cost of reducing the level of detail.

3.2 Image segmentation

Often the area to be investigated in the video is only a very small part of the total captured area. In the case of recordings by surveillance cameras, the interest often lies additionally on the time stamp of the video, which is usually located in the upper area. To extract only the relevant information, non-essential areas of the image can be removed and the regions needed for analysis can be selected. If necessary, relevant information can be reassembled into an image afterwards. This allows better focus on the essential content and simplifies purely cognitive analyses that occur later.

4 Image analysis

4.1 "Purely cognitive analyses"

The method behind pure cognitive analysis is called **correspondence analysis**. In correspondence analysis, the aim is to find out, on a purely visual basis, whether an unknown object in an image or video matches a known object. For the correspondence-based method, in addition to the image of the unknown object, at least one separate image is needed that depicts either the same object or a very similar object. To compare the objects in both images, remarkable structures, which describe the examined object, have to be determined. These noticeable structures, also called features, representing characteristics for comparing the objects. If the same features are identified in matching locations when comparing the objects, a match can be assumed. When selecting the features, it should be checked previously whether the structures are not image artifacts, which can arise, for example, from an underlying compression process. This would falsify a correspondence analysis. To prevent this, if objects from video data are compared or more than one comparison image of the object is available, correspondence features can be tracked over several frames. A disadvantage of the cognitive analysis is that it is inherently subjective in nature due to the selection of features. In order to find corresponding features, an appropriate level of detail must also be available. Accordingly, this method is also limited by the quality of the image and video material.

4.2 "Analysis of raw pixel values"

Basically, a digital image is divided into a limited number of elements, called pixels, where each pixel has a unique, fixed location in the image and is assigned one or more finite, discrete values [CD11]. Mathematically, a digital image (I) is usually described as a matrix (1)

$$I = i(x, y), \tag{1}$$

wherein it is divided into image rows (L) with indexes x (x = 0, 1, 2 ... L-1) and image columns (R) with indexes y (y = 0, 1, 2, ..., R-1). The pixels already described above, also called picture element, each represent an element in I. i(x, y) consequently represents the value of a picture element at its corresponding location coordinates. [NI20]

These values can now be examined with respect to forensic problems and provide information about various properties of image objects. This is the case, for example, with **colour analysis**. For this purpose, the pixel values of an image area are averaged and the values are analysed. We have observed that a transformation of the colour values into the HSV colour space ([CD11]) is better suited for colour analysis.

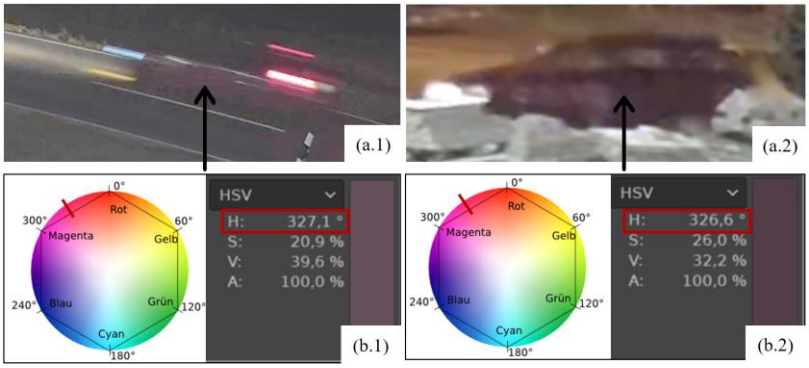

Fig. 4 Colour analysis of two vehicles captured by surveillance cameras of a car dealership near the crime scene (a). Both vehicles were observed to have a colour between magenta and red, which is illustrated by the colour circles and the red lines (b).

However, the lighting situation can have a great influence on the captured colour values and thus lead to irreversible shifts of the values in the colour space. In addition, if an object is further away from the camera, the number of available pixels is reduced, which means that all colours within the available pixels, for the object, are represented mixed. A similar phenomenon can be observed when working with highly compressed images, if several pixels are encoded and stored together during compression. In that case, an accurate colour analysis can be difficult or even impossible if the object to be analysed is represented by only a few pixels.

4.3 Extraction of semantic information

Machine learning, among other techniques, is used to extract semantic information from the images. For example, free AI frameworks can be used to extract and match faces.

To be able to analyse complete humans, another framework has been established, the neural network based "OpenPose". This is used to predict key points on the human body based on 2D image information. "OpenPose" is the first real-time multi-person system that can simultaneously recognize human body, hand, face and foot "keypoints" on individual images. The process is divided into several steps. After receiving a 2D image as input, the network generates a 2D confidence map for each body part position. In addition, 2D vector fields for part affinity fields (PAFs) are computed. These represent the directed connection from each keypoint to the adjacent keypoints of the same person. Now, based on the confidence maps, all graphs connecting different, possibly adjacency keypoints are created. Then, weaker links are removed using the PFAs. As a final step, the 2D positions of the anatomical keypoints for each person in the image are predicted from the information obtained in the previous steps. [CA18]

This framework is particularly useful in combination with externally generated

information. Thus, this can be used to match the motion or skeletal apparatus of a known person with an unknown person in the video. To do this, a digital skeleton (Rig) first has to be derived, also based on "OpenPose", from the underlying image information, which is obtained from individual images of a photogrammetric scan. Rigging as a working technique in 3D animation fundamentally refers to the construction of a skeleton or rig based on bones to determine how individual parts of a 3D model can be moved and can be applied analogously to a 2D context. By merging and comparing this "OpenPose"-Rig of a known person with the information obtained through "OpenPose" of the keypoints of an unknown person in an image or video, a conclusion can be made about whether the unknown person is the known person or not. In order to enable such a comparison, not only the known person himself, but also the surrounding areas as well as the crime scene must be captured photogrammetrically and transferred into a 3D reference model. In the resulting 3D space, measurements and comparisons of the rigs can be performed.

Fig. 5 Comparison of the digital skeleton of a known person, generated by "OpenPose", during an extended measurement service (a) with a digital skeleton of an unknown person, captured by a surveillance camera of a gas station (b).

This procedure builds on the assumption that the skeleton and thus also the digital skeleton of a person is unique [BE22].

The problem with all these methods is the usually very poor quality of the video and image material. This can lead to deviations in the predictions. In addition, the detection of faces seems to have difficulties in face recognition when other strong patterns dominate in the face area, which additionally obscures parts of the face, as well as when the level of detail of the face is too low.

In conclusion, many methods shown in shows like CSI are based on real methods, but their competence is highly exaggerated. The methods described here represent a more realistic state of forensic image and video analysis.

5 Case studies

The process of applying the individual steps does not necessarily follow a strict pattern. The individual steps can be applied repetitively and sometimes in a different order.

All methods of Super Resolution explained can be compared on the basis of the first case to be discussed. This involved an attempted arson. Video footage from a car dealership surveillance camera, which happened to be near the crime scene, showed a person with a dog (Fig. 1). The task was to analyse both the person and the dog. As can be seen in Fig. 1 original frame, the person occupies only a small part at the edge of the field of view of the surveillance camera. In order to be able to perform a cognitive analysis, the first task was to extract the relevant area by image segmentation and to enlarge it according to Super Resolution methods (Fig. 1 (a)-(d)). In this way, for example, a size classification of the dog could be performed and the basic clothing of the person could be described. In addition to the analysis of the person with the dog, two vehicles should also be analysed, which were captured by two cameras of the car dealership while passing by (Fig. 4 (a)). Here, the assumption arose that they could be the same or similar vehicles. In the process of these analyses, the vehicles were examined for their colouration with the help of colour analysis (Fig. 4 (b)), once colour normalization has taken place. Both vehicles were observed to have a colour between magenta and red (Fig. 4 (b) red lines).

The second case to be considered is a preliminary investigation on suspicion of extortionate kidnapping from several apartments. The aim was to determine whether the passing transporter (Fig. 2 (a), (b)) recorded by a surveillance camera at a gas station, with apparent advertising spaces, was the comparison transporter provided by the police (Fig. 2 (c)). Again, the key image sections were first extracted using image segmentation. Then, six frames could be matched and aligned based on keypoints of the van. Based on the cumulated image, the extraction of registrable features for the correspondence analysis (Fig. 2 orange and red circles) could take place. Particularly noticeable were individual lettering and applications (such as handles) on the back and side of the transporter (Fig. 2 (b)). Once sufficient features have been extracted, a comparison can be made with the reference transporter (Fig. 2 (c)). Here, not only matching features are noted, but also deviating conspicuous features. Another very characteristic feature is a logo on the bonnet of the comparison transporter, which can be recognized from the high-resolution images provided (Fig. 3 (a.1)). At the same time, several frames could be extracted from a video of the surveillance camera, on which frame-spanning dark pixels can also be registered on the bonnet of the transporter (Fig. 3 (a.1)). After an initial extraction of the relevant areas (Fig. 3 (b)) and a separate adjustment of contrast and brightness of the frame section (Fig. 3 (c)), both extracted regions could be aligned and adjusted to each other through an information matching (Fig. 3 (d)).

For matching people, as in the case of a robbery series, "OpenPose" can be used. In this robbery series, a few suspects were available (Fig. 5 (a)) as well as a surveillance video of a gas station, which had recorded the robbery (Fig. 5 (b)). Extended measurements of the subjects were taken, with the goal of creating Rigs for all individuals with the help of "OpenPose" (Fig. 5 (a) color-coded lines). "OpenPose" was also able to generate the respective Rigs for the people that were captured by the surveillance camera based on the 2D information (Fig. 5 (b) colour-coded lines). After creating a 3D model of the crime scene based on a photogrammetric scan, the Rigs could be matched. In this way,

an assignment of the known suspects to the persons recorded during the crime could take place.

Bibliography

[At13] Atanassov, K. et al.: Temporal image stacking for noise reduction and dynamic range improvement. In: Kennedy S. et al. (Hg.): Multimedia Content and Mobile Devices. IS&T/SPIE Electronic Imaging, Burlingame, California, USA, 2013

[BE22] Becker, S. et al.: COMBI: Artificial intelligence for computer-based forensic analysis of persons. In: KI - Künstliche Intelligenz, (in press) 2022

[BSS17] Bühler, P.; Schlaich, P.; Sinner, D.: Digitale Fotografie. Fotografische Gestaltung - Optik - Kameratechnik. Springer, Berlin, Heidelberg, Germany, 2017

[CA18] Cao, Z. et al.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018.

[CA22] Cardinale, F.; et al.: ISR. https://github.com/idealo/image-super-resolution. accessed: 31/05/2022

[CD11] Chanda, B.; Dutta Majumder, D.: Digital image processing and analysis. Second edition. PHI Learning Private Limited. Delhi, 2011.

[FA14] Fadnavis, S.: Image Interpolation Techniques in Digital Image Processing: An Overview. In: Int. Journal of Engineering Research and Applications (10), pp. 70–73, 2014

[GR90] Grimm, F.: Expertensysteme für den Einsatz von Subroutinenpaketen. Am Beispiel eines Expertensystems für Bildverarbeitung. Springer, Wiesbaden, Germany, 1990

[IQ90] Iqbal, K.: Enhancing the low quality images using Unsupervised Colour Correction Method. In: IEEE International Conference on Systems, Man and Cybernetics (SMC), Istanbul, Turkey. pp. 1703–1709, 2010

[KN19] Knop, D.: Schärfentiefe nach Maß. In: Biol. Unserer Zeit 49 (1), pp. 48–57, 2019

[LS17] Labudde, D.; Spranger, M.: Forensik in der digitalen Welt. Moderne Methoden der forensischen Fallarbeit in der digitalen und digitalisierten realen Welt. Springer, Berlin, Heidelberg, Germany, 2017

[NI20] Nischwitz, A. et al.: Bildverarbeitung. 4. Auflage. Springer, Wiesbaden, Heidelberg, Germany, 2020

[PMG13] Patel, V.; Mistree, K.; Gopalbhai, C.: A Review on Different Image Interpolation Techniques for Image Enhancement. In: International Journal of Emerging Technology and Advanced Engineering 3 (12), pp. 129–133, 2013

[PO05] Podlas, K.: The CSI Effect": Exposing the Media Myth. In: Fordham Intellectual Property, Media & Entertainment Law Journal 16, pp. 429, 2005