FORMS-Locks: A Dataset for the Evaluation of Similarity Measures for Forensic Toolmark Images

Manuel Keglevic and Robert Sablatnig
Computer Vision Lab, TU Wien
Favoritenstr. 9/183-2, A-1040 Vienna, Austria
mkeglevic@caa.tuwien.ac.at

Abstract

We present a toolmark dataset created using lock cylinders seized during criminal investigations of break-ins. A total number of 197 cylinders from 48 linked criminal cases were photographed under a comparison microscope used by forensic experts for toolmark comparisons. In order to allow an assessment of the influence of different lighting conditions, all images were captured using a ring light with 11 different lighting settings. Further, matching image regions in the toolmark images were manually annotated. In addition to the annotated toolmark images and the annotation tool, extracted toolmark patches are provided for training and testing to allow a quantitative comparison of the performance of different similarity measures. Finally, results from an evaluation using a publicly available state-of-the-art image descriptor based on deep learning are presented to provide a baseline for future publications.

1. Introduction

Lock snapping is a common way for forced entry in Europe. The unique imprints of the adjustable wrenches used for these break-ins significantly support the investigation of such offenses and are crucial as evidence in the following court cases. In Figure 1 for example, a snapped lock cylinder is shown. On the left side, the mounting point which is used to attach the lock inside the door is highlighted. This is the weak spot of the lock which snaps when enough leverage is applied. On the right side, pointing to the outside of the door, multiple overlapping toolmarks are visible.

One goal is the retrieval of toolmarks made by the same tool to link crimes together. However, manual examination of these toolmarks in order to find multiple uses of the same tool by forensic experts is a time consuming task due to the amount of samples. Therefore, an automatic filtering of the samples in order to reduce the amount of images requiring manual examination is desirable. Even though the development of automatic tools for the comparative examination of toolmarks has been in focus of the forensic community - since the validity of manual examinations by forensic experts has been challenged in court - only one dataset has been made publicly available so far. Most papers have been published with focus on obtaining statistical support for the notion of the *uniqueness* of toolmark patterns [6] to validate the identification of matching toolmarks as forensic evidence in court with the automatic comparison of toolmarks as a means to this goal. For instance, Bachrach *et al.* [1] used toolmarks made with 10 different screwdrivers of the same manufacturer and model number to examine the statistical distributions of similarity values. Petraco et al. [5] used 36 different screwdrivers, Spotts et al. [6] 50 sequentially manufactured slip-joint pliers, and Baiker et al. [2] 50 off-the-shelf screwdrivers of two different models. Only the NFI Toolmark dataset created by Baiker et al. [2] was made publicly available.

Figure 1: Snapped lock cylinder with the broken mounting highlighted on the left and the toolmarks on the right.

Further, in all those experiments, the toolmarks were created under constrained laboratory conditions. The tools and surface materials were hand-selected, the toolmarks were

made in a reproducible way using a fixed angle of attack, the lighting conditions were constrained, and the images or 3D surface scans are available in very high resolution; more than 400 pixels/mm in the case of the NFI Toolmark dataset. Even though we recently showed, using the NFI Toolmark dataset, that Convolutional Neural Networks (CNNs) can be successfully applied to learn a similarity measure for striated toolmarks [4], an assessment of the real-world performance of the automatic comparison of toolmarks is not possible without a new dataset.

Therefore, we collaborated with the forensic experts of the Criminal Intelligence Service Austria and the Austrian Police as part of the FORMS (Forensic Marks Search) project to create a dataset of toolmarks on lock cylinders which were seized during investigations of break-ins. The goal of this dataset is to cover their current use case: new lock cylinders are examined under a comparison microscope and an overview image in 10× magnification, which contains the whole toolmark, is taken and archived. Similar toolmarks are then found in a two step process. First, the overview images digitally stored in the archive are compared manually. Secondly, when a potential match is found, the actual cylinders are retrieved and compared in 20× magnification under the comparison microscope. In case the match is confirmed by the expert, an image of the aligned matching parts of the toolmarks is saved as evidence for court.

Since the camera attached to the microscope has a restricted resolution of 5MP the striated patterns of the toolmarks are not always visible. This shifts the focus of this dataset to the matching of impression marks left by the edge of the tool. Further, due to influence of the lighting conditions on the visibility of toolmark features, all images were captured under 11 different lighting settings to allow a quantitative assessment of their significance.

To permit the comparison of local image similarities, matching points on the toolmarks were marked by hand using a ground truth annotation tool. Similarly to the Photo-Tourism dataset [7], patches and matching and non-matching pairs are made available to allow a quantitative performance comparison. Additionally, the original images, manual annotations, and the annotation tool are provided.

In Section 2 the creation of the dataset and the GT annotation tool are described. In Section 3 the dataset with three different partitionings and the file format of the annotations are shown. Finally, in Section 4 an evaluation is performed using a state-of-the-art approach for comparing image similarities to provide a baseline for future publications.

2. Creation

In this section, first the image acquisition process for the dataset is described in detail. Since our dataset is motivated by the needs of forensic experts, it is based on the current workflow at the Austrian Police. Secondly, the annotation of matching points in the toolmark images and the tool developed for this purpose are presented.

Figure 2: Leica comparison microscope which is was used for capturing the toolmark images.

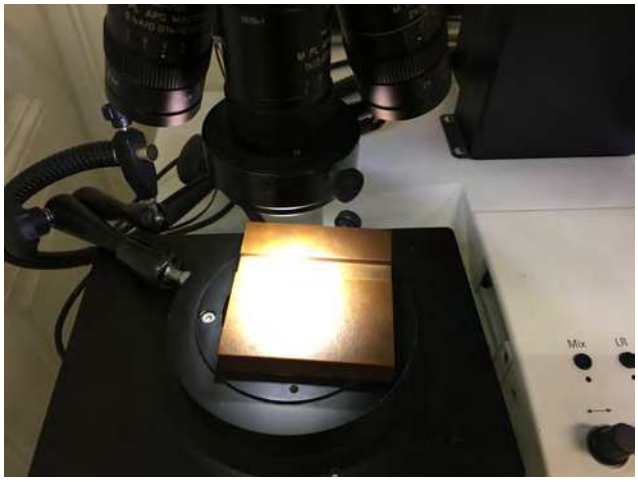

Figure 3: Closeup of the holding plate of the comparison microscope. The notch in the plate guarantees that the locks are inserted upright and the surface is flat.

2.1. Image Acquisition

For the creation of the dataset lock cylinders seized by the Austrian Police in the course of break-in investigations in Vienna during the years 2015 and 2016 were used. The comparison of toolmarks is conducted by the forensic experts using a Leica comparison microscope with lenses of varying magnification factor and an attached digital camera with a resolution of 5MP. In Figure 2 the comparison miroscope used is depicted and Figure 3 shows a closeup of the holding plate where the lock cylinders are placed. To

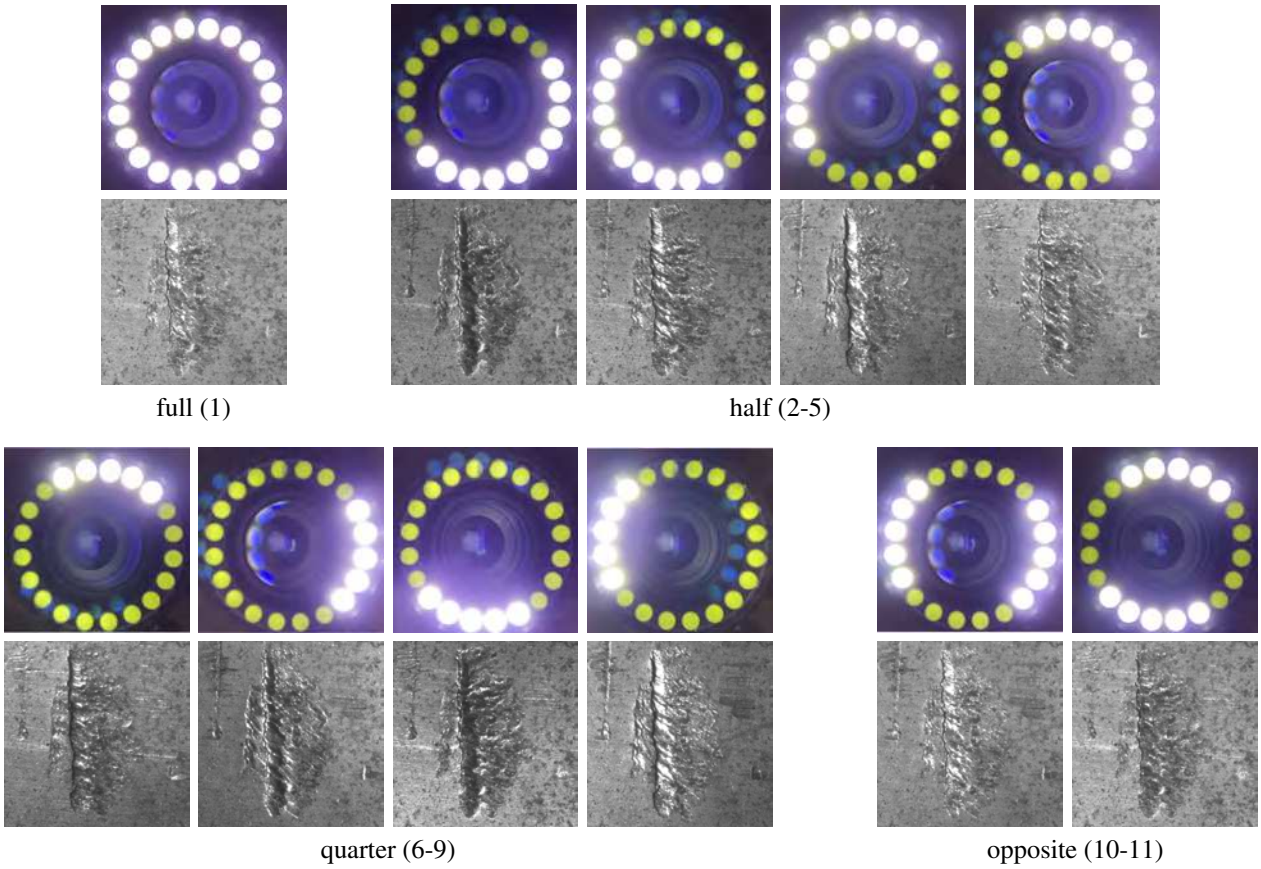

Figure 4: The 11 different lighting settings of the ring light used which are organized into 4 different groups. To illustrate the influence on the toolmark images, corresponding crops are shown below.

vary the lighting, an adjustable ring light with 11 different settings and a flexible spotlight are available. This way, the contrast of certain toolmark features can be enhanced. In Figure 4 the different settings of the ring light are depicted. Further, using the toolmark image crops the influence on the visible toolmark patterns is illustrated.

During a normal workflow, new lock cylinders are first cataloged by using a 10× magnification lense to create an overview image containing the whole toolmark; or toolmarks in case multiple marks are present. Since both jaws of the adjustable wrenches have independent patterns, both sides of the cylinders are captured. By always placing the cylinders upright with the broken (inner) part of the lock facing left, all toolmark images can be compared without rotating or flipping the images. We used a similar strategy to create the dataset. Even though striated toolmarks are better visible under 20× magnification, it would not be possible to caputure the whole toolmarks with one image. This would on the one hand require a stitching of multiple images which in turn introduce artifacts at the borders. On the other hand, it would complicate and lengthen the capturing process significantly and therefore contradict our initial objective to alter the workflow of the forensic experts as little as possible. Moreover, since one goal of this dataset is to investigate the influence of different lighting conditions and the robustness of a similarity measure to variations in lighting, we captured each side of the cylinder with all 11 available lighting settings. These settings, which are divided into 4 different groups, are shown in Figure 4. In each group different fractions of the ring light are lit up and in each group (except *full* illumination) the direction of the light can be changed. This information is made available by coding the lighting condition into the image names, *i.e.* images with filenames ending with "01" belong to group *full*, "02","03","04","05" belong to group *half*, and so on. For images of cylinders from the year 2015 the file ending also indicates the exact lighting direction, *e.g.* "06" indicates the group *quarter* and direction from the top. Due to an issue with the light ring, for the images of cylinders from the year 2016 only the group can be derived. In Table 1 the number of tools, locks and captured images in total are shown for each year.

As single toolmarks without matching counterparts cannot be used for evaluation and training, we focused on the

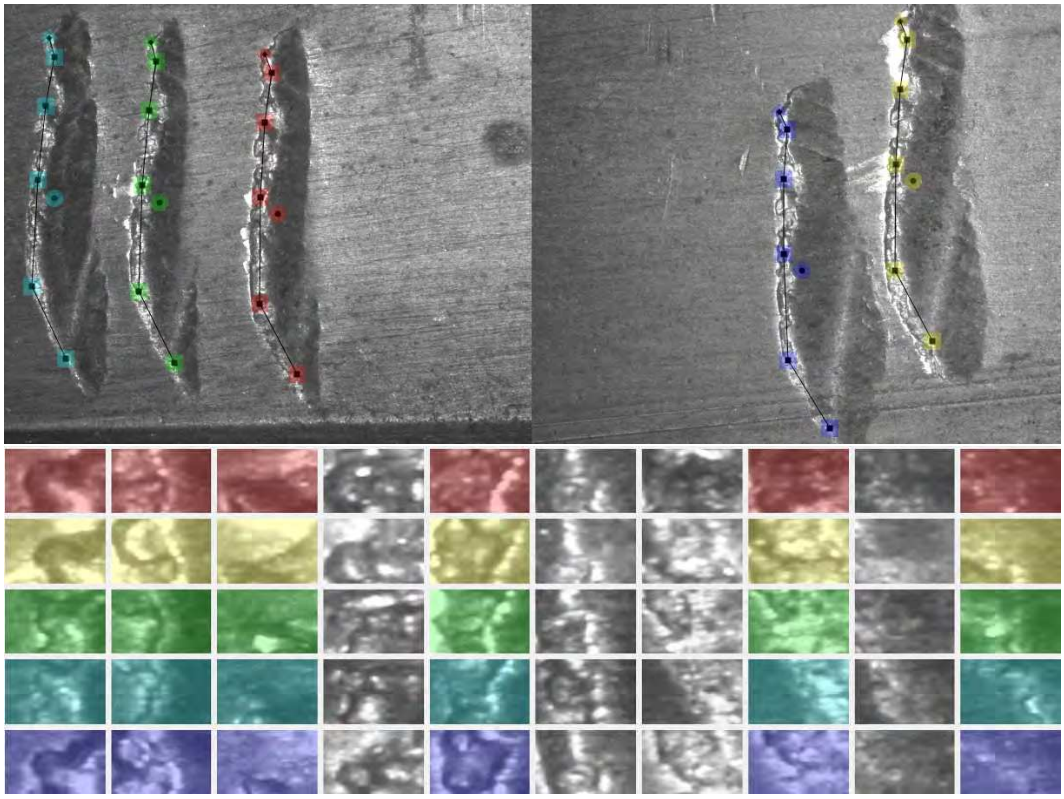

Figure 5: Annotation tool which assists the manual drawing, cloning and fitting of polylines. The merged images, which show multiple toolmarks made by the same tool, are displayed at the top. For each toolmark, a drawn polyline has been cloned and manually fitted to its edge using translations and rotations. Patches are extracted along these polylines to aid the user at precisely aligning the polylines. This is shown at the bottom. Control points on a polyline and corresponding patches have the same color.

	Tools	Locks	Sides	Images
2015	25	115	230	1,782
2016	23	82	164	1,263
total	48	197	394	3,046

Table 1: Statistics of the captured toolmark images divided by year.

lock cylinders which have been previously identified as matching by the forensic experts. This restricts the total number of different tools in the dataset to the 48 which have been used in linked criminal cases, i.e. crime series. In total 3,046 images were captured.

2.2. Ground Truth Annotation

Even though the toolmark images acquired in the previous section are already annotated with the used tool, only 48 distinct tools with 96 distinct jaws, *i.e.* jaws, are available. To provide a more fine-grained annotation of the images, matching image regions, i.e. matching patches, are desired. A manual annotation of matching patches in the toolmark images is not feasible since hundreds of patches per image have to be matched. Yet, the toolmark images provide advantageous properties to simplify this task. Since our regions of interest lie on the edges of the toolmarks, a polygonal chain (polyline) can be used to describe the points on this edge parametrically. Further, the same polyline can be used to describe the edges in matching toolmark images; albeit transformations are required to fit a polyline to a new image. The transformations necessary are given by the image acquisition process and the properties of the lock cylinders and jaws of the adjustable wrenches. First, the area of interest containing the toolmarks on a cylinder lock is approximately a flat surface and the capturing angle is orthogonal to this surface. Therefore no perspective transformations are necessary. Further, the distance of the camera to the surface is always the same since the used lenses have a fixed focus and therefore focusing of the image is performed by moving the lock cylinder into focus. Thus, also scaling transformations do not have to be considerd. By restricting the allowed transformations to translation and rotation, the

polylines can be efficiently fitted to the edges of matching toolmarks.

We implemented this approach as a plugin for the image viewer nomacs. Similarly to nomacs, the so called PatchMatchingPlugin is open source and available on gitub[1]. This tool allows the user to draw polylines along the edges of the toolmarks. Further, the polylines can be cloned and manually fitted to a matching toolmark in the same image using rotations and translations. The resulting polylines and their clones then define matching points along their line segments. Matching patches can be extracted by choosing one point on a polyline and using the transformation matrices to map this point to a clone of this polyline. In order to assist the annotation process, the patches on a polyline and its clones can be displayed with varying patch size and distance between the points. For the annotation, matching toolmarks on two different cylinders and multiple distinct toolmarks on one cylinder are considered. Since the PatchMatchingPlugin only allows the annotation of a single image, first, for both sides of each cylinder the image in which the toolmark is best visible is chosen by the user. Thereafter, matching images are merged into one image which is then used for the annotation process.

Figure 5 shows the annotation result on an example. At the top, the merged image is shown. In this case, three distinct marks are visible in the image on the left side and two on the right side. By drawing a polyline, first a toolmark edge is marked; in this case the red one. Then, for each distinct toolmark a clone is created and manually fitted. At the bottom, extracted patches along the polyline and corresponding points on the clones are shown to help the user adjust the fitting. Color coding is used to associate the clones with their respective patches. This is done by coloring the control points of the polyline differently for each clone. The interpolated points between the control points are shown in gray. Finally, the polyline and the transformation matrices for the clones are stored in a JSON (JavaScript Object Notation) file. Since the lock cylinders were not moved during the image acquisition, the same annotations can be used for the images of all lighting settings.

The annotation process was performed for all 96 merged images in the dataset and the results were verified by a forensic expert. The example depicted in Figure 5 can be considered a best-case annotation result. Depending on the hardness and shininess of the lock cylinder material and the force used by the intruder, the fitting of the different clones may be significantly worse. Especially, in case a toolmark imprint is not deep enough, reference points which are crucial to aligning the clones, like the start and end of a toolmark or distinct patterns, may not be clearly visible. Further, in case overlapping toolmarks are present, finding a clear consistent toolmark is challenging.

[1] https://github.com/nomacs/nomacs-plugins

3. Dataset

In this section, the two ways the dataset[2] is provided are decribed. First, the 96 merged images for each jaw of a tool with their respective annotations and 11 lighting settings, *i.e.* 1,056 images and 1,056 JSON files. Secondly, to allow a comparable evaluation of different similarity measures without the need of a patch extraction, image patches are provided with a list of 100,000 matching and non-matching pairs. For this, three distinct partitionings focusing on different challenges of the dataset are presented.

3.1. Annotated Images

On average 2.9 toolmark images were combined to create the 96 merged images for each jaw of a tool. In Table 2 the detailed distribution is depicted. In most cases, only two toolmark images per side are available. However, in some cases as much as 10 different toolmark images of the same tool could be combined. In a few cases no toolmarks were visible on one side of the lock cylinder and therefore no image was captured.

	Matching Toolmark Images (per side)									
	1	2	3	4	5	6	7	8	9	10
2015	3	20	8	11	4	1	0	1	0	2
2016	8	22	8	4	5	2	0	0	0	0
total	11	42	16	15	9	3	0	1	0	2

Table 2: table

An annotated image consists of one merged image of a minimum dimension of 2592x1944 pixel and a JSON file of the same name with the prefix “patches.json”. The images are concatenated by placing them side by side. Therefore, the width of the merged images is a variable multiple of 2592 pixels. The filename contains a yearly index for the tool, “15” or “16” for the year, “1” or “2” for the side of a tool, and “01”-“11” indicating the lighting configuration as depicted in Figure 4. For example, the image showing toolmarks of the second side of the tool with index 1 in the year 2016 captured under light settings 3 (group *half*) is named “1_16_2_03.png”. In Table 3 the number of polylines and clones in the dataset are shown. A merged image can contain multiple polylines to describe partial toolmarks and none if no toolmarks are visible.

	Merged Images	Polylines	Clones
2015	50	52	175
2016	46	44	115
total	96	96	190

Table 3: Number of polylines and clones in the dataset.

[2] https://www.caa.tuwien.ac.at/cvl/forms-locks/

Listing 1: Examplary JSON annotation with one polygon described by two control points and four clones of this polyline defined by their transformation matrices.

```
[{
  "polygon": {
    "points": [
      [3807.3409391715777,
        779.96621476630935],
      [3813.4372570232063,
        1334.7311392645361]
      ]
  },
  "clones": [{
    "transform": [
      [1,0,0],
      [0,1,0],
      [0,0,1]
    ]
  }, {
    "transform": [
      [1,0,0],
      [0,1,0],
      [2689.1673532866389,
        -103.6374034776909,1]
    ]
  }, {
    "transform": [
      [1,0,0],
      [0,1,0],
      [5011.6001511800332,
        -253.80000832021267,1]
    ]
  }, {
    "transform": [
      [1,0,0],
      [0,1,0],
      [-2618.6000989574227,
        145.80000477969634,1]
    ]
  }]
}]
```

The annotated polylines and clones are stored in a JSON file. An example is shown in Listing 1. The top level structure is an array containing one or multiple entries with polylines (called "polygon") and clones. The polylines are defined by their control points with an array of 2-dimensional image coordinates. For each clone a 3x3 transformation matrix is given which allows the mapping of the points on the polylines to the actual image coordinates.

3.2. Extracted Image Patches

In order to allow a comparable evaluation of methods for local toolmark similarities, a traningset and testset of extracted patches is provided. Since the number of tools for the years 2015 and 2016 are approximately the same, the sets are divided by year. However, as shown in Table 1 more images are available for the year 2015. Further, Table 2 indicates that the merged images for the year 2015 contain more images on average which is crucial to providing matching patches for training. Therefore, the trainingset is created with images of the year 2015 whereas the patches for the testset are extracted from images of the year 2016.

As described in section 2.2, each location on a clone in the dataset is defined by the position on the polyline, i.e. the distance from the first control point calculated by following the line segments, and the transformation matrix for the clone.

The extraction of patches is then performed as follows: for each clone in each merged image, 64×64 patches are extracted along the polyline with a stepsize $k = 64$ and $k = 8$ for the testset and trainingset, respectively. Each patch in a dataset can therefore be uniquely identified by the merged image with specific lighting configuration L, the polygon index I_p, the location on the polygon t_i indexed in k steps, and the clone index I_c. The filename of the patches are then composed by a counting number, index of the tool, year, side, "fg" for foreground, I_p, t_i, I_c, and L. For example, "025795_30_15_2_fg_01_0011_02_03.png" shows the patch #25795, extracted from the merged image "30_15_2_03.png" on the 11th position on the first polygon, on clone 2 and with lighting configuration 3. Only patches on the polylines are extracted, patches in the background are ignored. Three distinct partitionings of the dataset FORMS-Locks, FORMS-Locks-RR, and FORMS-Lock-Lighting are proposed:

FORMS-Locks The focus of this partitioning lies in finding matching patches without considering their orientations. Therefore, the orientation of each patch is fixed to the rotation of transformation matrix of the corresponding clone. This way, matching patches are guaranteed to be oriented the same. In this partitioning, patches are defined as matching if they are extracted from matching positions of clones in any lighting setting.

FORMS-Locks-RR This partitioning is similar to FORMS-Locks, however the robustness in regard to variations in orientation is evaluated additionally. For this, patches are extracted with random orientations for the testset. For the trainingset, at each location 10 patches with random orientations are extracted. This increases the number of the patches in the trainingset 10 fold, compared

to FORMS-Locks. Matching patches are equally defined as in FORMS-Locks.

FORMS-Locks-Lighting-RR The goal of this partitioning is to isolate the influence of varying lighting conditions on the performance. Therefore, to remove errors introduced by the manual annotation of the matching toolmarks and variations due to cylinder lock materials and force applied, matching points on clones are ignored and only patches from exactly the same image location but with different lighting settings are considered as matching patches. Similarly to FORMS-Locks-RR, the patches are extracted with random orientation.

In Table 4 the number of patches in each partitioning is shown. The lists with 50,000 randomly sampled matching and non-matching pairs from the testsets for evaluation are provided as CSV files which include two patch indices and a "0"/"1" indicator in each line for matching and non-matching pairings, respectively. Further, in addition to the 64×64 sized patches, scaled-down 32×32 patches are provided for all partitionings.

	#Patches	
	train	test
FORMS-Locks	41,030	25,014
FORMS-Locks-RR	410,300	25,014
FORMS-Locks-Lighting-RR	410,300	25,014

Table 4: Number of patches in the datasets.

4. Evaluation

In order to provide a baseline, the PN-Net published by Balntas is evaluated on the dataset *et al*. [3]. The proposed CNN implements a so-called triplet architecture to learn an embedding in which the L_2 distance defines a similarity measure. The source code for training and evaluation of PN-Net is freely available on github[3]. Evaluated on the Photo-Tour dataset [7], which consists of matching and non-matching 32×32 images patches extracted from 3D mapped tourist photos with three different subsets (Liberty, Notredame, and Yosemite), they achieve state-of-the-art results with a false positive rate at 95% recall (FPR95) of 4-10%; depending on the subsets used for training and evaluation. For each partitioning, we trained the PN-Net on the 32×32 image patches of the trainingset and evaluated the image pairs provided with the testsets. Similarily to [3] the FPR95 is used as performance metric since it allows an intuitive assessment of the number of expected false positives in case almost all the true positives are correctly identified, *i.e*. 95%. Further, it enables a comparison with results on the Photo-Tourism dataset and thus an estimation of how *hard* the proposed dataset is.

[3] https://github.com/vbalnt/pnnet

	Balntas *et al*. [3]
FORMS-Locks	78,77%
FORMS-Locks-RR	83,24%
FORMS-Locks-Lighting-RR	31,68%

Table 5: Results FPR95.

As shown in Table 5 the best results are achieved on FORMS-Locks-Lighting-RR with an FPR95 of 31,68%. Even though this result is still far worse than on the Photo-Tourism dataset, this indicates, that adapting to the various lighting conditions is the least challenging problem. The results on the FORMS-Locks and FORMS-Locks-Lighting-RR are far worse with an FPR95 of 78,77% and 83,24%, respectively. However, the difference between these two partitionings is only about 4-5%. This shows, that the CNN does not simply learn to distinguish different orientations of the patches and the most challenging problem in the dataset is actually matching patches from different toolmarks.

5. Conclusion

In this paper we presented a new toolmark dataset based on real break-ins investigated by the Austrian Police. Since no similar dataset exists yet, this contribution is crucial for the development of methods for the automatic comparison of toolmark images. We extensively described how the dataset was created and manually annotated. In addition to the 3,046 captured images and annotations describing matching points in these images, the annotation tool itself is made publicly available. Further, three different partitionings, with more than 25,000 patches in the testset, are provided to allow quantitative comparisons. Finally, a baseline evaluation using a state-of-the-art CNN architecture is presented which shows that computing similarities for forensic toolmark images is an open research topic with great room for improvement.

For future work, a new annotation of the dataset for the automatic localization of the toolmark edges and an even bigger dataset which allows a shift of focus from image patches to complete toolmark images is planned.

Acknowledgements

This work has been funded by the Austrian security research programme KIRAS of the Federal Ministry for Transport, Innovation and Technology (bmvit) under Grant 850193. We would like to thank the forensic experts of the Criminal Intelligence Service Austria and the LKA Wien (AB08 KPU) for their help. The Titan X used for this research was donated by the NVIDIA Corporation. This work was supported by *die Buben*.

References

[1] B. Bachrach, A. Jain, S. Jung, and R. D. Koons. A Statistical Validation of the Individuality and Repeatability of Striated Tool Marks: Screwdrivers and Tongue and Groove Pliers. *Journal of Forensic Sciences*, 55(2):348–357, 2010. 1

[2] M. Baiker, I. Keereweer, R. Pieterman, E. Vermeij, J. van der Weerd, and P. Zoon. Quantitative comparison of striated toolmarks. *Forensic Science International*, 242:186–199, 2014. 1

[3] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk. PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors. *ArXiv*, 2016. 7

[4] M. Keglevic and R. Sablatnig. Learning a Similarity Measure for Striated Toolmarks using Convolutional Neural Networks. In *Proceedings of the 7th IET International Conference on Imaging for Crime Detection and Prevention (ICDP)*, 2016. 2

[5] N. D. K. Petraco, H. Chan, P. R. D. Forest, P. Diaczuk, C. Gambino, J. Hamby, F. L. Kammerman, W. Brooke, T. A. Kubic, L. Kuo, G. Petillo, E. W. Phelps, A. Pizzola, and D. K. Purcell. Application of Machine Learning to Toolmarks - Statistically Based Methods for Impression Pattern Comparisons. Technical report, NCJRS (239048), 2012. 1

[6] R. Spotts, L. S. Chumbley, L. Ekstrand, S. Zhang, and J. Kreiser. Optimization of a Statistical Algorithm for Objective Comparison of Toolmarks. *Journal of Forensic Sciences*, 60(2):303–314, 2015. 1

[7] S. A. J. Winder and M. Brown. Learning Local Image Descriptors. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007. 2, 7