Automated Detection and Categorization of Genital Injuries Using Digital Colposcopy

Kelwin Fernandes[1,2](✉), Jaime S. Cardoso[1,2], and Birgitte Schmidt Astrup[3]

[1] INESC TEC, Porto, Portugal
{kafc,jaime.cardoso}@inesctec.pt
[2] Universidade do Porto, Porto, Portugal
[3] Institute of Forensic Medicine, University of Southern Denmark, Odense, Denmark

Abstract. Despite the existence of patterns able to discriminate between consensual and non-consensual intercourse, the relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries. The testimony of the physicians when assessing these lesions has been questioned in court due to several factors (e.g. a lack of comprehensive knowledge of lesions, wide spectrum of background area, among others). Thereby, it is relevant to provide automated tools to support the decision process in an objective manner. In this work, we compare traditional handcrafted features and deep learning techniques in the automated processing of colposcopic images for genital injury detection. Positive results where achieved by both paradigms in segmentation and classification subtasks, being traditional and deep models the best strategy for each subtask type respectively.

Keywords: Genital injury · Digital colposcopy · Deep learning · Handcrafted features · Image processing

1 Introduction

The relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries [1–3]. Since genital lesions are frequent in both, consensual and non-consensual intercourse [1,4], the existence of a pattern of genital injury able to discriminate trauma seen in rape cases and trauma seen following consensual sexual intercourse has been a matter of study in the past [3]. Typical different patterns were analyzed by several authors [3,5]. Slaughter et al. [5] suggested that multiple genital lesions at multiple locations are frequent in rape victims, while single lesions in the posterior forchette are predominant in consensual sexual intercourse. Astrup et al. [3] suggested a higher frequency of abrasions, haematomas and multiple lesions in rape cases. Also, Astrup et al. [3] confirmed a higher frequency of lesions in locations other than the 6 o'clock position and the presence of larger and more complex lesions in non-consensual cases.

Although the existence of such pattern has been validated by several studies, the debate continues. Legal experts suggest the lack of comprehensive knowledge

L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 251–258, 2017.
DOI: 10.1007/978-3-319-58838-4_28

of lesions sustained during consensual sexual intercourse as a key problem [1]. Moreover, the expert responsible for conducting such evaluations as well as the physical analysis in sexual assault victims itself differ around the world [2]. For instance, in the US most examinations are done by specially trained nurses, while in many European countries the examinations are performed by gynecologists [6]. Other countries like Denmark delegate this responsibility to forensic pathologists [2]. Given the wide spectrum of background knowledge of experts and the low inter-evaluator agreement [4], the expert testimony given by these professionals in cases of genital lesions in sexual assaults has been questioned in court in several countries [2].

In this work we propose a preliminary framework for the automated detection and categorization of genital injuries on digital colposcopies using image processing and machine learning techniques. Although these techniques have been successfully used in other medical applications in order to improve and support the medical decision process (e.g. [7]), even on digital colposcopy analysis [8,9], this is the first attempt to address the detection of genital injuries from a computational perspective. Building an objective data-driven system, able to provide a unified framework for the analysis of genital lesions in digital colposcopy may increase the reliability of genital trauma findings in legal rape complaints from both, medical teams and legal experts. From a technical point of view, we study the impact of handcrafted and deep-features in the automation of several tasks of the proposed system.

2 Preliminary Definitions

In this section, we describe common concepts that are fundamental in this work. Specifically, the type of lesions of interest and the investigative techniques used in their detection. Further details about these concepts can be found in the medical literature [10].

2.1 Investigative Techniques

The usual methodologies for the detection of genital injuries cover the naked eye inspection, the colposcope and inspection after application of toluidine blue dye (see Fig. 1) [10]. In this work we merely include the two latter which allow automation.

- **Colposcope:** The investigator inspects the external genitalia and afterwards the vagina and cervix using digital colposcope. A colposcope is a binocular instrument that magnifies and illuminates a given inspected area.
- **Toluidine Blue Dye:** After inspection, a blue dye is applied to the genital mucous membranes and then wiped off. Toluidine blue stains exposed cellular nuclei but not intact mucosa, thus enhancing areas of surface disruption.

(a) Colposcope with light source (b) Colposcope (c) Toluidine blue dye

Fig. 1. Examples of images from several acquisition techniques

(a) Laceration (b) Abrasion (c) Hematoma

Fig. 2. Examples of genital injury on digital colposcopy

2.2 Injuries

The European and Australian categorization of genital injuries is used in this work, which comprises laceration, abrasion and hematoma (see Fig. 2) [10].

- **Laceration:** Discontinuity of epidermis and dermis. Caused by blunt force such as tearing, crushing, or overstretching.
- **Abrasion:** Traumatic exposure of lower epidermis or upper dermis. Most often caused by lateral rubbing or sliding against the skin in a tangential manner. The outermost layer of skin is scraped away from the deeper layers.
- **Contusion/Hematoma/Bruise:** Traumatic extravasation of blood in tissues below an intact epidermis. Caused by blunt force.

3 Proposed Methodology

We subdivide the whole framework into a set of five predictive tasks: light source detection, segmentation of gloves, detection of the investigative technique (colposcope and toluidine blue dye), segmentation of the toluidine blue dye stained regions, classification of lesions and discrimination of consensual and non-consensual intercourse.

3.1 Deep Learning Strategies

Two deep learning paradigms were used in this work: domain-specific neural networks and pre-trained architectures with additional fine-tuning.

Domain-Specific Neural Networks. In this case, networks are trained from scratch using our domain specific dataset. The architecture used in this case was common for all the subtasks. We used a standard convolutional neural network (CNN) [11] with an alternating sequence of Convolutional and Max Pooling layers, followed by a Dropout layer and ending with a sequence of dense layers. We validated several activation functions for dense layers obtaining the best trade-off between convergence speed and performance with leaky rectifier units [12]. The shape and activation function of the final output layer depends on the task type. For classification tasks with global output (i.e. one output per image), we used a layer with N output values and soft-max activation function, being N the number of categories. On the other hand, for segmentation tasks, the last layer has size $rows \times cols$, returning an activation value for each output pixel. This approach has been used as an efficient alternative to encoder-decoder networks [13] traditionally used for segmentation [14]. The final parametrization for each task was fine-tuned using an independent validation set.

Pre-trained Neural Networks. Here, we used pre-trained architectures (i.e. VGG16, VGG19, ResNet50, Inception V3) on the ImageNet dataset. Then, the last dense layer of each network is fine-tuned using our own data. This approach was only applied to the classification sub-tasks. The best network was chosen using the validation data.

3.2 Strategy Based on Hand-Crafted Features

In this section we describe the methodology used in the corresponding subtasks using hand-crafted features.

Light Source Detection. Since the presence of artificial light is spatially located in approximately the same round area (see Fig. 1a), we simplified the task of segmenting the lighted-unlighted areas to a global binary classification task. The traditional pipeline designed for this task works as follows. First, the specular reflections (SR) are removed and the images are quantized using an index of K colors obtained with the K-means algorithm [15]. Then, the image is segmented into a group of nested circles. The feature vector used to fit the binary classification model is the concatenation of the relative frequency difference for each color inside and outside the circles. Also, we include the average intensity difference inside and outside each circle. The binary classifier is chosen using K-fold cross-validation from a set of estimators, namely, Random Forest, AdaBoost, Support Vector Machine, Gradient Boosting and Logistic Regression.

Gloves and Toluidine Blue Dye Segmentation. For the segmentation of gloves and regions stained with toluidine blue dye we used a common framework based on superpixels classification. Images are segmented into a fixed number of regions using SLIC Superpixels (i.e. clustering on the color-distance space).

Table 1. Features used in the segmentation of gloves and toluidine blue dye.

Category	Description
Texture	Local Binary Patterns histogram. Number of edged pixels (using Canny edge detector)
Color	Relative frequency of each quantized color (color index built using K-means). μ and σ of each channel in the RGB and HSV color spaces
SR	Number of blobs and relative size
Shape	Basic statistics (i.e. min, max, μ, σ) of the pixel locations to the image center, borders and ROI mask. Size

Then, for each segment we extract a broad set of features (see Table 1) and train a binary classifier to label each cluster. The best classifier is chosen in the same way than in the previous section. Finally, images are post-processed to remove small blobs. For the toluidine blue dye, we also include an additional binary classification task regarding the presence or not of the staining substance in the image. For this subtask, we used the features from the categories texture, color and SR (see Table 1) on the whole image.

Lesion Classification and Consensual vs Non-consensual Classification. Manually computing high-level hand-crafted features for the classification of these two tasks is a extremely challenging and time consuming labor. Thereby, we decided to apply standard pipelines for the classification of images using state-of-the-art image descriptors (e.g. SIFT, SURF). In this sense, after computing the keypoints and their corresponding descriptors, a Bag-of-Visual-Words using the K-Means algorithm on the descriptor space is trained. Then, each image is encoded by the normalized histogram of vocabulary words found on it. Finally, a binary classifier is trained using the image's encoding. The best binary classifier is chosen using K-fold cross validation in a similar way than in the previous sub-tasks. We concatenated the features obtained with the SIFT and SURF algorithms to build the binary classification model.

4 Experiments

In the experimental assessment of the proposed methods a dataset with 394 images collected by the *Southern Denmark Sexual Assault Referral Centre* was used (78 images from non-consensual intercourse and 316 from consensual intercourse). For further details about the acquisition process refer to the original source [4]. We divided the database into three disjoint sets for training, validation and test using a standard 60-20-20 partition. The distribution of images with and without artificial light, with each color of gloves and from consensual and non-consensual cases was kept constant. Since the frequency of hematoma in our dataset was very limited, abrasion and hematoma cases were combined as a single class in the categorization of lesions. The performance of each classification strategy was measured using accuracy (Acc) and macro-averaged area under the ROC curve (AUC). In counterpart, since the proposed deep architectures generate fuzzy segmentations, the fuzzy Dice coefficient is used. The fuzzy Dice coefficient (f-Dice) can be defined in a straightforward manner using the fuzzy definitions of set intersection and set cardinality.

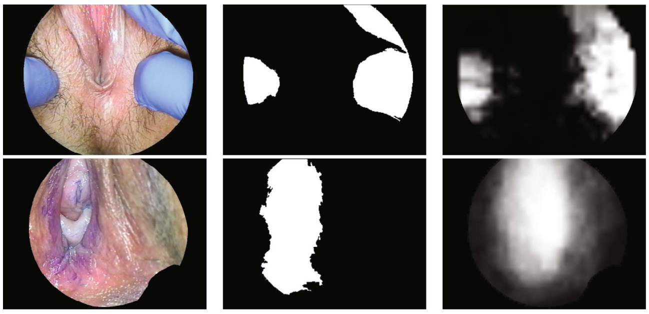

Fig. 3. Results obtained by the gloves (**top**) and toluidine blue dye (**bottom**) segmentation strategies. **Left:** original image. **Middle:** Handcrafted features. **Right:** Deep Learning. (Color figure online)

Table 2a and b show the results for the classification and segmentation subtasks respectively. The results are overall satisfactory, being able to provide positive predictive results for all the proposed subtask. As observed in related areas, deep learning strategies performed better than traditional pipelines in most classification tasks in term of AUC. However, models based on hand-crafted features

Table 2. Performance of the traditional and deep strategies on each subtask.

(a) Classification subtasks

Sub-task	**Hand-crafted**		**Deep**		**Pre-Trained**	
	Acc	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**
Light source	**100.00**	**100.00**	98.73	**100.00**	97.47	99.75
Toluidine blue dye	93.67	97.44	**96.20**	**99.16**	89.87	98.51
Lesion detection (binary)	**68.35**	62.71	25.32	56.10	67.09	**71.44**
Lesion categorization	72.15	**63.47**	**74.68**	61.67	65.82	63.34
Consensual/Non-consensual	**84.81**	79.46	81.01	77.68	**84.81**	**88.10**

(b) Regression subtasks

Sub-task	**Hand-crafted**	**Deep**
Gloves	**84.99**	63.40
Toluidine blue dye	**84.30**	70.66

achieved better performance than deep architectures trained from scratch in general. Regarding accuracy, traditional models surpassed deep learning strategies in several cases, even when achieving low AUC values. These behaviors are probably due to the high complexity of deep models whose parameters cannot be properly estimated using small datasets. The low performance of the lesion detection and categorization tasks when compared to the consensual/non-consensual problem suggests that the models might be overfitting to features unrelated with the genital injuries. Thereby, further validation is required to ensure that both groups where handled indistinguishably.

The landscape is different for the segmentation subtasks, where traditional techniques outperformed deep strategies by a large margin. As can be observed in Fig. 3, the boundaries obtained by the traditional strategies are clearly defined, while the ones obtained by the deep learning methods are very smooth. This may be due to the subsampling effect generated by the convolutional/max-pooling layers or due to the high number of parameters involved in these models, turning the fitting of sound models a challenging task with small datasets. Since our segmentation subtasks are used as pre-processing steps to remove undesired regions from the input image, sharp boundaries are preferred. However, smooth behaviors are desirable in plenty medical applications and might be useful in the spatial location of genital injuries.

5 Conclusions

Despite the existence of patterns able to discriminate between consensual and non-consensual intercourse have been proved, the relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries. Being the lack of comprehensive knowledge of lesions a driving factor in the acceptance of this type of evidence in courts, it is fundamental to provide objective methods to support the expert's decision. In this work, we proposed a framework that covers the preliminary steps in the automated detection of genital injuries on digital colposcopies using computer vision and machine learning. We compared the performance of traditional pipelines with handcrafted features and deep learning approaches in several subtasks.

We validated that, for our problem, both strategies are complementary being the former more suitable for segmentation tasks and problems with easily transmitted high-level discriminative information (e.g. light detection) and the later more suitable for complex tasks, where building high-level informative features is challenging (e.g. lesion classification). However, deep learning strategies require a lot of data to be trained, achieving low performance in some tasks when trained from scratch. In the future, we will work in the spatial detection of lesions in order to suggest areas of interest in the image to the physicians. Also, we will explore hybrid architectures, able to combine the best features from each paradigm.

Acknowledgements. This work was funded by the Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/93012/2013.

References

1. Astrup, B.S., Lauritsen, J., Ravn, P., Thomsen, J.L.: Genital lesions after consensual sexual intercourse: they are frequent and they last for several days. In: 19th World Meeting of the International Association of Forensic Sciences (2011)
2. Astrup, B.S., Ravn, P., Lauritsen, J., Thomsen, J.L.: Nature, frequency and duration of genital lesions after consensual sexual intercourse–implications for legal proceedings. Forensic Sci. Int. **219**(1), 50–56 (2012)
3. Astrup, B.S., Ravn, P., Thomsen, J.L., Lauritsen, J.: Patterned genital injury in cases of rape–a case–control study. J. Forensic Legal Med. **20**(5), 525–529 (2013)
4. Astrup, B.S., Lauritsen, J., Thomsen, J.L., Ravn, P.: Colposcopic photography of genital injury following sexual intercourse in adults. Forensic Sci. Med. Pathol. **9**(1), 24–30 (2013)
5. Slaughter, L., Brown, C.R., Crowley, S., Peck, R.: Patterns of genital injury in female sexual assault victims. Am. J. Obstet. Gynecol. **176**(3), 609–616 (1997)
6. Payne-James, J., Busuttil, A., Smock, W.: Forensic Medicine: Clinical and Pathological Aspects. Cambridge University Press, Cambridge (2003)
7. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imaging Graph. **31**(4), 198–211 (2007)
8. Fernandes, K., Cardoso, J.S., Fernandes, J.: Temporal segmentation of digital colposcopies. In: Paredes, R., Cardoso, J.S., Pardo, X.M. (eds.) IbPRIA 2015. LNCS, vol. 9117, pp. 262–271. Springer, Cham (2015). doi:10.1007/978-3-319-19390-8_30
9. Huang, X., Wang, W., Xue, Z., Antani, S., Long, L.R., Jeronimo, J.: Tissue classification using cluster features for lesion detection in digital cervigrams. In: Medical Imaging, International Society for Optics and Photonics, p. 69141Z (2008)
10. Astrup, B.S., Lykkebo, A.W.: Post-coital genital injury in healthy women: a review. Clin. Anat. **28**(3), 331–338 (2015)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint. arXiv:1505.00853 (2015)
13. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint. arXiv:1511.00561 (2015)
14. Pan, J., McGuinness, K., Sayrol, E., O'Connor, N., Giro-i Nieto, X.: Shallow and deep convolutional networks for saliency prediction. arXiv preprint. arXiv:1603.00845 (2016)
15. Kasuga, H., Yamamoto, H., Okamoto, M.: Color quantization using the fast k-means algorithm. Syst. Comput. Jpn. **31**(8), 33–40 (2000)