A Transfer Learning Approach for the Tattoo Detection Problem

Rodrigo Tchalski da Silva, Heitor Silvério Lopes
Graduate Program on Electrical Engineering and Informatics (CPGEI)
Federal University of Technology Paraná (UTFPR), Curitiba, Brazil
rodrigo.tds@gmail.com, *hslopes@utfpr.edu.br*

***Abstract*—Tattoos are still poorly explored as a biometrics factor for human identification, especially in public security, where tattoos can play an important role for identifying criminals and victims. Tattoos are considered a soft biometrics, since they are not permanent and can change along time, differently from hard biometrics traits (fingerprint, iris, DNA, etc). The identification of tattoos are not simple, since they do not have a definite pattern or location. This fact increases the complexity of developing models to address this problem. In addition, the tattoo identification roadmap is very complex, including several steps and, in each step, specific methods need to be developed. Among the several problems identified in this roadmap, we tacked the identification problem, which is defined as: given an image of a person, determine if there is a tattoo or not. We present a deep learning model based on transfer learning for the tattoo detection problem. We also used data augmentation to improve the diversity of the training sets so as to achieve better classification accuracy. Along the work two new datasets for tattoo detection were created. Several comparative experiments were done to evaluate the diversity of images in the datasets, and the accuracy of the proposed model. Results were very promising, achieving an accuracy of 95.1% in the test set, and a F1-score of 0.79 in an external dataset. Overall, results were satisfactory, given the complexity of the problem. Future work will focus on expanding the datasets created and addressing the other problems of the tattoo roadmap.**

***Index Terms*—Tattoo detection, Transfer learning, Deep neural network, Pattern recognition**

I. Introduction

Tattoos are not only an expression of art and a kind of "customization" of the human body. They are also considered as biometric identifications that, through their almost unique features, can be used as a form of people identification.

Similar to clothing, tattoos are classified as soft biometrics, that is, something that can be modified and is not permanent [1]. However, tattoos are a human identification factor that is very relevant from the applied point of view. Therefore, the development of studies to improve the quality of the identification systems based on this biometrics has significant relevance not only to applied sciences but, also, to security.

Its application in public security, in particular, is of great importance, since the problem of identifying individuals is a fundamental issue for this area. The identification of suspects or criminals, and the recognition of victims in disasters are two examples of the use of tattoos as a biometrics. In these cases, tattoos be helpful to public security agents in cases where the hard biometrics (digital, iris, face, hand palm, etc.) may not be available, as in the case of image capture on surveillance cameras.

Hard biometrics have well-defined patterns and localization. Therefore, the process of identifying these patterns is more precise. However, tattoos do not have patterns, and may have any shape, color, size or location on the body. Therefore, one of the great challenges is, given an image (of a person), devise if there is a tattoo or not. More specifically, the tattoo is the only region of interest, and it can be considered the foreground and all the remaining of the image is considered the background. This is the tattoo detection problem.

The main objective of this work is to present a method for detecting tattoos in images. This is the first issue of the tattoo recognition roadmap, to be presented later. This study is based on applying the transfer learning method to extract features from images and, then, to detect tattoos using a trained binary classifier. For this, different well-known deep learning networks were tested in the transfer learning step. Along the work, it was needed to create a high-quality dataset for training the models. Specific data augmentation procedures were also used to improve the robustness of the trained classifiers.

II. Tattoo Recognition Roadmap

Tattoo recognition is a complex process that involves several tasks and, to the best of our knowledge, no model was found in the literature capable of addressing all the problems related in this roadmap. After compiling many studies carried out in this area, we found that tattoo recognition is divided into two main steps: pre-processing and recognition (Fig. 1).

The pre-processing step is responsible for preparing the original image for the recognition step. It includes the following problems:

- Detection: determines whether an image (of a human) has a tattoo or not.
- Location: finds where in the image the tattoo is found, and returns a bounding-box around the corresponding region of the image.
- Segmentation: crops out the exact contour of the tattoo from the rest of the image, removing all the surrounding background.
- Semantic Segmentation: crops further the tattoo image separating each distinct object represented in the tattoo.

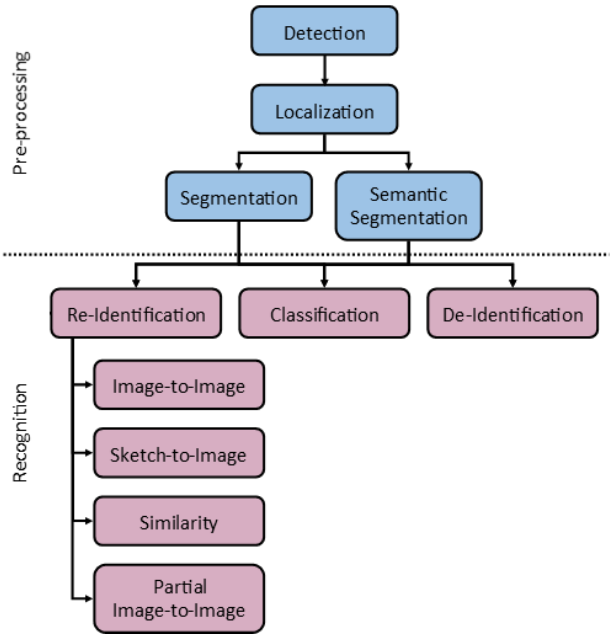

Figure 1. Tattoo recognition roadmap

The pre-processing tasks have the challenge to prepare the image to be submitted to the next step, receiving a raw image and returning an image without the noise represented by the background. This process is as important as the recognition process, because the better the pre-processing, the more efficient the recognition.

The recognition processing, which are the methods that effectively perform the tattoo recognition, includes the following problems:

- Re-identification: searches for an image in a dataset of tattoo images, returning the one that is the most similar to the searched tattoo. This process can be further divided into other specific models:
  - Image-to-Image: consists of, given a sample tattoo image, finding the most similar image in a database.
  - Sketch-to-Image: consists of, given a hand-drawn sketch of a tattoo, finding the most similar image in a database.
  - Similar Groups: consists of, given a sample tattoo image, finding a group of tattoos that are similar to the one searched and that have the same pattern, but not necessarily the original image searched.
  - Partial Image-to-Image: consists of, given a partially occluded tattoo image, finding images that have the searched image as part of a complete image in the database.
- Classification: given a tattoo image, give a description for the object or objects that make up that image, returning labels to the input image.
- De-Identification: consist of a process of erasing the tattoo from an image, a process also known as anonymization.

As mentioned before, the tattoo detection problem consists in determining whether an image contains a tattoo or not. Figure 2 shows some examples of images of people with and without tattoos. Despite the theoretical simplicity of the concept, the detection process is not such a simple task at all, as there are no defined standards for what a tattoo is in terms of patterns of shape, color, size, proportion to the individual and, mainly, its location on the body. In addition, a single image can have several tattoos. Furthermore, the background where the image of a person was captured can introduce significant noise to the detection process, since its complexity may be confused with tattoos. Some of these issues are shown in the images shown in Fig. 2, for both tattoo and non-tattoo images.

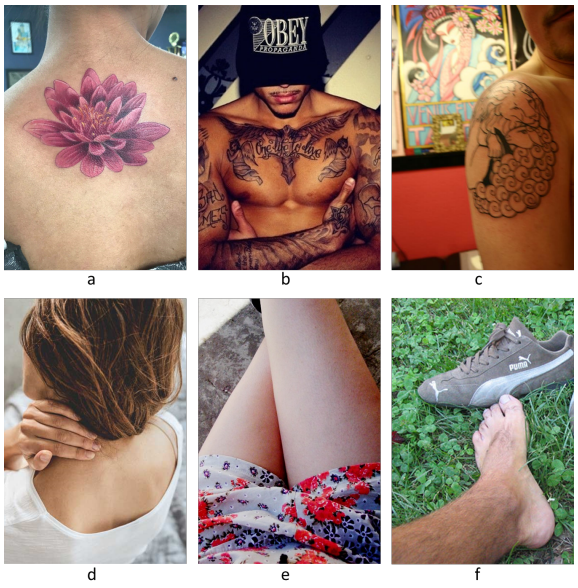

Figure 2. Example of images, used for the detection task, of people with and without tattoos. Image "a" is an example of well-behaved situation, where the tattoo is a well-defined, and the surrounding background is a clear skin. The Image "b" there are multiple tattoos, and a bounding-box of them will include other areas of the body where there is no tattoo. Also, the letters of the hat may disturb the recognition process. In image "c" the painting in the backward has similar patterns to the tattoo, and they are close to each other. The image "d" is a good example of non-tattoo, easily identified. In the image "e", the design of the clothes in contrast with the skin may be confused as a tattoo. In image "f" the green background contrasts with the remaining foreground, and the design of the tennis may be confused as a tattoo.

III. Literature Review

Tattoo detection plays a fundamental role in the initial image filtering and data selection, and its importance has been neglected for many years by the scientific community [2], and the first publication on this topic came up only recently, in 2015 [3].

After that seminal publication, several other studies followed. Table I presents a summary of the results recently published for the tattoo detection problem. In this table, column “Best Result 1” refers to results obtained when using the same dataset for both, training the model and testing the model, evaluating its accuracy. In the column “Best Result 2”, the accuracy presented was achieved by testing the model

Table I
TATTOO DETECTION PUBLISHED RESULTS

Ref.	Year	Method	Best Result 1	Best Result 2
[4]	2015	not cited	96.30% acc.	-
[2]	2016	CNN	98.80% acc.	93.78%
[5]	2016	AlexNet + 2-Class SVM	99.83% acc.	-
[6]	2017	AlexNet + 2-Class SVM	99.83% acc.	-
[7]	2016	Decision tree	52.38% acc.	-
[8]	2016	Faster R-CNN	98.25% acc.	80.66%
[9]	2019	Faster R-CNN	87.10% recall (WebTattoo) 61.70% recall (Tatt-C)	80.00%

using a different dataset from that used to train the model (this issue will be addressed later in this paper).

In Table I it is noticed a diversity of machine learning methods as well as seemingly good results. However, training and testing over the same dataset may lead to biased results, since generalization capability of the classifiers are not evaluated. This fact raises an important issue, the datasets used. Table II describes the databases used in each study. Due to the large diversity of tattoo types and the lack of standards for capturing images, datasets can be very different to each other. As a consequence, it is difficult or even unreasonable to compare results.

It is important to notice that the first published results appeared in response to the challenge published by NIST [3]. In this scenario, four institutions presented results, with the company MorphoTrek presenting the best performance with 96.3% of accuracy [4]. Unfortunately, the algorithms used by NIST participants were not published. This fact turned out impossible to carry out external validation tests, which was criticized in [2].

Table II
TATTOO DATASETS REFERENCED IN THE TATTOO DETECTION BIBLIOGRAPHY.

Ref.	Dataset Name	# Tattoo Images	# Non-Tattoo Images
[4]	Tatt-C	1,349	1,000
[2]	Tatt-C	1,349	1,000
	NTU_Flickr	5,740	4.260
[5], [6]	Tatt-C	1,349	1,000
[7]	Unidentified	547	-
[8]	Tatt-C	3,839	-
	PASCAL Visual Object Classes (VOC) 2007	-	9,963
	NTU_Flickr	5,740	4,260
[9]	Tatt-C	1,349	1,000
	NTU_Flickr	5,740	4.260
	WebTattoo	300,000	-

Based on this scenario, [2] suggested to evaluate whether the dataset available in [3] and [4] was sufficiently comprehensive to ensure that the results presented could be generalized. Therefore, the authors presented a Conventional Neural Network (CNN) trained in two scenarios: the Tatt-C dataset [3] and a dataset obtained from Flickr (NTU_Flickr). The experiment consisted of training the network with one of the datasets and validating with the other, and vice-versa. Initially, in the same scenario presented in [4], the CNN described in [2] had a slightly higher performance, increasing the previous accuracy from 96.3% to 98.8%. In subsequent tests, networks trained on the NIST dataset and validated on the NTU_Flickr dataset performed less well than the other way around. Finally, the authors showed that as the training dataset increases, the result accuracy also improves.

In [5] and [6], authors also propose using a CNN for tattoo detection, also basing their study on the Tatt-C dataset. The proposed model consists in extracting features through fine tuning the AlexNet network and, then, applying a linear Support Vector Machine (SVM) to determine whether an image has tattoos or not. The proposed algorithm was also compared with [4], and obtained an accuracy of 99.83%, that is, an improvement of 3.2% compared to the best initial result.

Another approach based on decision trees was presented by [7], this time in its own dataset, with less expressive results, reaching only 52.38% of accuracy.

In [8] the authors present a deep learning region-based method, the Faster R-CNN, which is based on a fine-tuning of the VGG_CNN_M_1024 network. The training data was also based on Tatt-C dataset, but now joining other 9,963 images without tattoos divided in 20 object categories from the PASCAL Visual Object Classes (VOC) 2007 dataset [10]. Their results were also compared with those presented in [4], and its performance had an accuracy of 98.25%, meaning 1.95% better than that presented by MorphoTrek in [4].

More recently, [9] presented a detection model also using a Faster R-CNN. In this model, the detection problem was classified as an instance of the image recovery system, where learning and detection were performed simultaneously. The authors also present a result based on the recall percentage, which was compared with the results obtained in [8]. While [8] presented a recall of 45% to 0.1 FPPI (false positive per image) for the Tatt-C dataset, the authors in [9] presented a result of 61.7% for the same dataset and 87.1% to a dataset obtained from the internet (called WebTattoo). In summary, it is difficult to compare different works due to differences in test procedures, metrics and the datasets used.

IV. DATASETS

Datasets are a fundamental part for all machine learning methods. In the one hand, choosing the correct dataset to perform a study is directly related to the quality of results. On the other hand, looking at the results without evaluating the dataset used can lead to misconceptions about the real quality of results. Considering that a dataset is a sample of the real world, it is particularly important for the efficiency

of machine learning methods that the datasets used reflect the same diversity. Also, for multi-class datasets, the balance of samples in the classes is another important issue since, in general, classifiers are strongly biased towards the majority class. Unfortunately, many real-world datasets do not follow such principles. For instance, the dataset Tatt-C presented in [3], widely used in the literature, have images of the non-tattoo class predominantly of faces. Possibly, this can bias a classified trained with this dataset, acquiring a misconcept that images without tattoos are generally those with faces. This issue was criticized by [2] in their publication. Despite this, that database was the most used, to date, for studies involving tattoos.

Taking into account the previous considerations, we designed datasets for this specific study taking care to maximize the image diversity and minimize possible biases in the results. Considering the need of two classes, namely, tattoo and non-tattoo, both must have diversity not only in terms of the specific part of the body that is in focus but, also, in the amount of background, distance to the tatoo within the image and framing pattern.

In addition, it is desirable to obtain data from different sources, in order to avoid possible bias due to the source of the information. In the next Section, experiments will show how this issue was addressed.

Two datasets were created, namely, TattDetectB and TattDetectF, with images extracted from internet at Bing[1] and Flickr[2], respectively. Each dataset was composed of 2,000 images of people, 1,000 for the tattoo and 1,000 for the non-tattoo class.To obtain images for the proposed dataset, a web scraping technique was used. It consists of scanning internet pages, identifying images, and capturing. A Python script was used to perform web scrapping. For each website we performed the web scraping searching for images with and without tattoos separately. Also, aiming at improving the diversity of images, we searched for images with tattoos combined with specific parts of the body, such as back, shoulder, arms, legs, etc. Such a procedure was done for both, TattDetectB and TattDetectF and, also, for tattoos and non-tattoos classes. Overall, this procedure helped to provide a good balance within the datasets.

For the two datasets, the same criterion was used for both, tattoo and non-tattoo classes, including only images containing at least one person in the image, or a part of a person. Differently from other datasets in the literature, no other random image (without a person) was included..

Here it is important to note that the TattDetectF dataset used in this study is not the same as the NTU_Flickr dataset mentioned in the studies carried out by other authors, already cited here and presented in Table II. Although the data was scrapped from the same source, they are different.

To date, we did not find any specific methodology for comparing tattoo datasets, regarding their diversity and complexity, so as to compare our proposed datasets with other ones.

[1]http://bing.com
[2]http://flickr.com

V. Methods

In this work we used the transfer learning technique, which has been shown excellent results in many classification problems, specially for image processing [11]–[13]. The basic idea is to use a CNN architecture, trained for a given problem, and re-use part of this architecture for other problem (Fig. 3). But, usually using the same type of data, in our case, images.

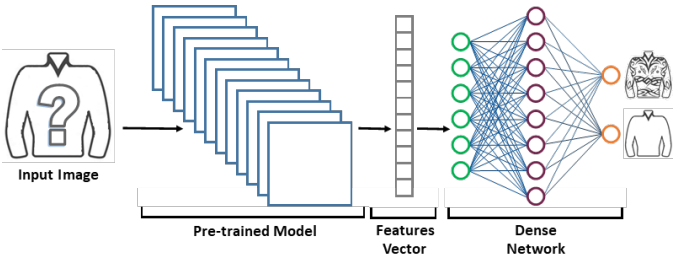

Figure 3. Proposed transfer-learning model

The last layers of the trained network is excluded and the remaining is re-used as a feature extractor. In other words, the knowledge learned by the trained network will be transferred to another similar problem. Therefore, using this procedure, a pre-trained network receives an image and provides a feature vector that represents that image in a high-dimensional embedded space.

In our working pipeline, the images of the training set are presented to the feature extractor and the output vector is forwarded to the input from a dense neural network, which is the trainable classifier for tattoo or non-tattoo classes.

We applied two different approaches by using the dense neural network. Firstly, using the same dataset for training and testing the model, with a 10-fold cross validation procedure. Secondly, using one dataset to train the model and a different dataset to test it. In both approaches the classifier was a dense network with a single hidden layer with 100 neurons, ReLu activator function, Adam solver, regularization of 0.0001 and 200 iterations to train the network.

When data is scarce or with low variety, a convenient way to expand data is by using data augmentation. With limited data, many problems may appear, such as overfitting and poor generalization capability. Those problems can be alleviated by using data augmentation methods, specific for the nature of the data being processed. In the case of images, there are many possible transformations [14]. In our experiments two scenarios using data augmentation were created. The first one, the data were augmented 6 times, and the second, augmented 12 times. For the first case, 6 randomly chosen transformations were applied, out of the 12 following ones: zoom, vertical mirroring, horizontal mirroring, rotation, warp perspective, Poisson random noise, Gaussian random noise, salt and pepper random noise, random contrast and brightness, Gaussian blur and a bilateral filter. For the second case, all the above-mentioned transformations were applied once each.

Also, we aimed to discover which CNN performed better as the feature extractor for our classification problem. Therefore,

the classification accuracy of tattoo detection was used for measuring the performance. The following architectures were tested as feature extractors: SqueezeNet [15], Inception-v3 [16], VGG-16 and VGG-19 [17].

VI. Experiments and Results

In this session, the results obtained for the experiments are reported and commented. The experiments were carried out using the Orange platform [18] and complemented with scripts in the Python language, both running in the Windows 64 bits environment, on a desktop computer with Intel i7-8565U CPU @1.80GHz processor, and 16 GB of memory.

A. Evaluation of feature extractors

In the first experiment, we aimed at answering the following question: "Which feature extractor can lead to better detection results ?". As mentioned before, we used SqueezeNet, Inception-v3, VGG-16, VGG-19. For these DNNs, the length of the feature vectors are 1,000, 2,408, 4,096 and 4,096, respectively.

In order to evaluate the quality of the feature extractors, they were input to a dense neural network (fully-connected) to classify each image into two classes: tattoo or non-tattoo. Both datasets created (TattDetectB and TattDetectF) were used, for training and testing with the four DNNs. We used the 10-fold cross-validation procedure and the average accuracy was reported. Results were quite similar to each other, since all feature extractors achieved very good results. Therefore, since the difference between the performances of the feature extractors was irrelevant, we elected Inception-v3 for the further experiments.

B. Evaluation of the effect of data augmentation

Since the former experiment did not indicate preference for a specific DNN architecture, the next experiment included all of them.

In the literature it is well-known that data augmentation is a valuable strategy for improving the quality of the classifier, especially for improving its generalization capability. Therefore, experiments were done to answer the question: "Does data augmentation applied to the training set improves the detection capability of the classifiers ?"

First, an "internal" baseline must be established. For this purpose, two experiments were done: first training with TattDetectB and testing with TattDetectF and, then, vice-versa. Results are shown in the left side of Table III.

Next, two new groups of datasets were created. The first, by augmenting each original image 5 times, generates a dataset with 6,000 images (the 1,000 original images plus 5,000 augmented images). The second, by augmenting 12 times each original image, generates a training dataset with 13,000 images. The same data augmentation procedures were applied to both, tattoo and non-tattoo images, of the original datasets. Therefore, the following new datasets were created: TattDetectB_Aug6, TattDetectF_Aug6, TattDetectB_Aug13, and TattDetectF_Aug13.

Table III
Results for the data augmentation tests.

Baseline			Augmented		
Training	Test	Accuracy	Training	Test	Accuracy
TattDetectB	TattDetectF	93.95%	TattDetectB_Aug6	TattDetectF	94,90%
			TattDetectB_Aug13	TattDetectF	95,10%
TattDetectF	TattDetectB	88.40%	TattDetectF_Aug6	TattDetectB	89,39%
			TattDetectF_Aug13	TattDetectB	90,24%

Now, using the new augmented datasets it is aimed to verify if data augmentation can improve upon the baseline results. That is, if the use of an augmented dataset increases the generalization ability of the classifier. Therefore, four experiments were run, and results are shown in the right side of Table III.

Comparing the accuracies of the baseline and the augmented experiments, it is inferred that data augmentation do improve the generalization capability of the classifier. The amount of improvement is small, possibly due to the fact that the baseline is high. Also, augmenting more (*_Aug13) the original dataset leaded to even better results. Comparatively, the classifier trained with TattDetectB_Aug13 was the best performing, and it will be used in the next experiments.

C. Comparison with other dataset

In order to perform an "external" evaluation of our approach, it would be desirable to compare its performance with other datasets published in the literature. However, the direct comparison with those works are not possible because some works published results by training and testing in the same dataset. In most cases, there is no information about how the dataset was split into training and testing datasets or if cross-validation was used instead. Also, most of the datasets used in previously published works are no longer available for downloading. We succeeded to find only one dataset with reasonable parameters for testing (two classes, relatively balanced, and a large number of images). Therefore, the question to be answered is: "How does our approach perform with an external dataset?".

The results for this experiment are presented in Table IV, where we first trained the network with our datasets and, then, tested with the NTU_Flickr dataset. Next, we trained with the NTU_Flickr dataset and tested with our datasets. Observe that the NTU_Flickr is unbalanced, with more tattoo images than non-tattoo images (see Table II). Consequently, accuracy is an inadequate performance measure, and the F1-score is used instead. For a fair comparison with our previous results (Table III), it is necessary to report the corresponding F1-score: when trained with TattDetectB_Aug13 and tested with TattDetectF, the F1-score was 0.95; and when trained with TattDetectF_Aug13 and tested with TattDetectB, the F1 score was 0.90.

Based on the results of Table III, we used both the classifiers trained on the augmented datasets, i.e., TattDetectB_Aug13

Table IV
RESULTS FOR THE CLASSIFICATION USING AN EXTERNAL DATASET.

Training	Test	F1-score
TattDetectB_Aug13	NTU_Flickr	0.78
TattDetectF_Aug13	NTU_Flickr	0.79
NTU_Flickr	TattDetectB	0.53
NTU_Flickr	TattDetectF	0.56

and TattDetectF_Aug13, to classify the NTU_Flickr dataset. Results are shown in Table IV.

When training with our datasets and testing with the NTU_Flickr, results were reasonable good, despite the results be less than 10% lower than those achieved in Table III. On the other hand, when training with NTU_Flickr and testing with our datasets, the results showed a larger drop in performance.

To investigate the possible reasons for these differences in performance, Figure 4 shows the confusion matrices for these experiments. In (a), results seems to be relatively balanced, although in (b) they are not, with many tattoo images being classified as non-tattoo. A visual inspection of the TattDetectF_Aug13 dataset, regarding tattoo images classified as non-tattoos, indicated that they were either very small tattoos or tattoos covering a large part of the body. This fact suggests that the TattDetectB_Aug13 dataset has a wider range of tattoo sizes in the images, compared with the TattDetectF_Aug13.

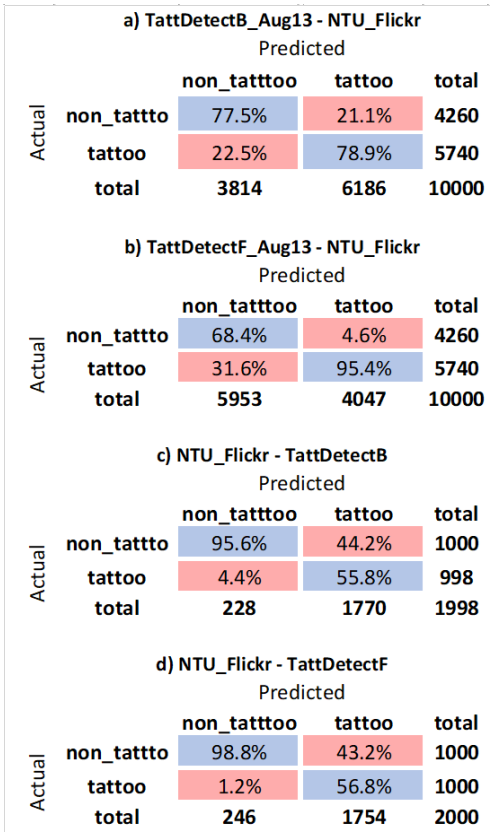

Figure 4. Confusion matrices for the experiments with an external dataset.

Observing the results of the confusion matrix in (c) and (d), a systematic unbalance was found. The network trained with NTU_Flickr dataset classified wrongly as tattoo 44.2% of the non-tattoos images of the TattDetectB dataset, and 43.2% of the TattDetectF dataset. The reasoning for this requires a visual inspection in the NTU_Flickr dataset. It was built using only human images in the tattoo class, and random images in the non-tattoo class, i.e., images of animals, sights, objects, drawings, cars, flowers, etc. On other hand, our datasets were built using only images of people and human body parts, with and without tattoos. This fact misleaded the classifier trained with the NTU_Flickr to classify anything human-like as the tattoo class.

D. Qualitative analysis

Finally, a qualitative analysis was carried out with the objective of verifying in which scenarios our approach had classification errors. Such analysis could shed a light into which kind of tattoos and non-tattoos are more difficult to be classified.

The confusion matrices from two of the best-performing train-test pairs shown in Table III are shown in Figure 5. On the one hand, when trained with TattDetectB_Aug13 and tested with TattDetectF more non-tattoos were wrongly classified as tattoos (8.47%) than the opposite (0.66%). On the other hand, when trained with TattDetectF_Aug13 and tested with TattDetectB more tattoos were wrongly classified as non-tattoos (16.19%) than the opposite (0.25%).

Although the results are obtained in terms of overall accuracy, this qualitative analysis helps us to realize that the datasets still deserve a little more attention, especially regarding to their image diversity, since this is, possibly, the cause of the asymmetry in the results reported above.

In the case of the model trained with the TattDetectB_Aug13 dataset, which had a larger error when classifying non-tattoos, the dataset was inspected. We observed that the non-tattoo part of the dataset is composed of many clean images, with a large amount of images with light backgrounds, and without much visual pollution. On the other hand, the wrongly classified non-tattoo images had more colorful backgrounds or people wearing more colorful clothes with details.

The same qualitative analysis was done with the TattDetectF_Aug13 dataset. We found that the wrongly classified images were those that had tattoos with less details, smaller in size to the image or less colorful, a class of images with less samples in trained dataset.

To exemplify the qualitative nature of this analysis, Fig. 6 brings some samples of images that were incorrectly classified as tattoos. It was possible to notice the following: elements with high contrast (a, b), colorful backgrounds (c, d, e, h), many details (h, i), people wearing colorful clothes (j, k), confused or blurred images (f), images with some colored element different from the rest of the image (g, k).

Similarly, Fig. 7 brings samples of images that had tattoos but were classified as non-tattoos. In general, errors were due to the small size of the tattoo, regarding the size of the image

a) TattDetectB_Aug13 - TattDetectF

Actual	Predicted: non_tatttoo	tattoo	total
non_tattto	99.34%	8.47%	1000
tattoo	0.66%	91.53%	1000
total	914	1086	2000

b) TattDetectF_Aug13 - TattDetectB

Actual	Predicted: non_tatttoo	tattoo	total
non_tattto	83.81%	0.25%	1000
tattoo	16.19%	99.75%	1000
total	1192	808	2000

Figure 5. Confusion matrices for different training and testing datasets.

Figure 6. Examples of non-tattoos wrongly classified as tattoos.

those in which the tattoos were hidden or very small (a, b, c, d, e, f), with many people (e, f), person full of tattoos (g, h) or unfocused (i).

VII. Conclusions

The present study aimed to present a model based on transfer learning applied to the problem of detecting tattoos in images, that is, given an image, to determine whether there is a tattoo on it or not.

This problem is the initial part of the tattoo identification roadmap, which involves a series of steps, each with its importance within the process (Figure 1).

From an applied point of view, and specifically in applications for public security, the use of tattoo recognition can significantly contribute to the work of identifying individuals and, thus, the development of new techniques that are robust has a great applied value.

The results presented in this research project showed to be significant, as it brought a new approach to the proposed problem as well as showing robustness in the results presented.

Regarding the feature extraction process, it was found that all DNN's tested had an good performance and similar results, reaching an average accuracy of 96.82% using a dense neural network as a classifier, with 10-fold cross-validation.

We also observed that data augmentation is effective in providing more robustness for the classification process. A small difference in performance (~5%) was observed using TattDetectB_Aug13 and TattDetectF_Aug13 for feature extraction, and TattDetectF and TattDetectB for testing, respectivelly. Such assimetry raised the need for a qualitative analysis.

We also applied our approach to classify another dataset found in the literature (NTU_Flickr [2], [8], [9]). In general, results were good, considering that the type of images of NTU_Flickr may be of different categories. Some imbalance in the results were observed, and they were due to the type of non-tattoo images of the NTU_Flickr dataset. Results suggested that the datasets proposed in this work are more realistic (than NTU_Flickr), keeping in mind that tattoo identification only makes sense in images of humans. It is possible that this fact contributed to reach better results.

However, it is important to emphasize that a quantitative comparison of classification performance with other works is not possible due to the methodological differences between works. It should be pointed, also, that the former studies with tattoos were based on the the Tatt-C dataset, provided by NIST, which was discontinued over time and is no longer found for download. Actually, the lack of standardized datasets for tattoo detection is a great drawback for this area of research.

Future work will include the expansion of our datasets, with increased diversity and quality, so that, once put at the public domain, we can foster more research in this area. Also, a deeper study about the effect of image properties, such as sizes, proportion of the tattoo in the image, illumination, effect of colors, effect of the complexity of the tattoos, etc, will be interesting research directions to be sought in the near future.

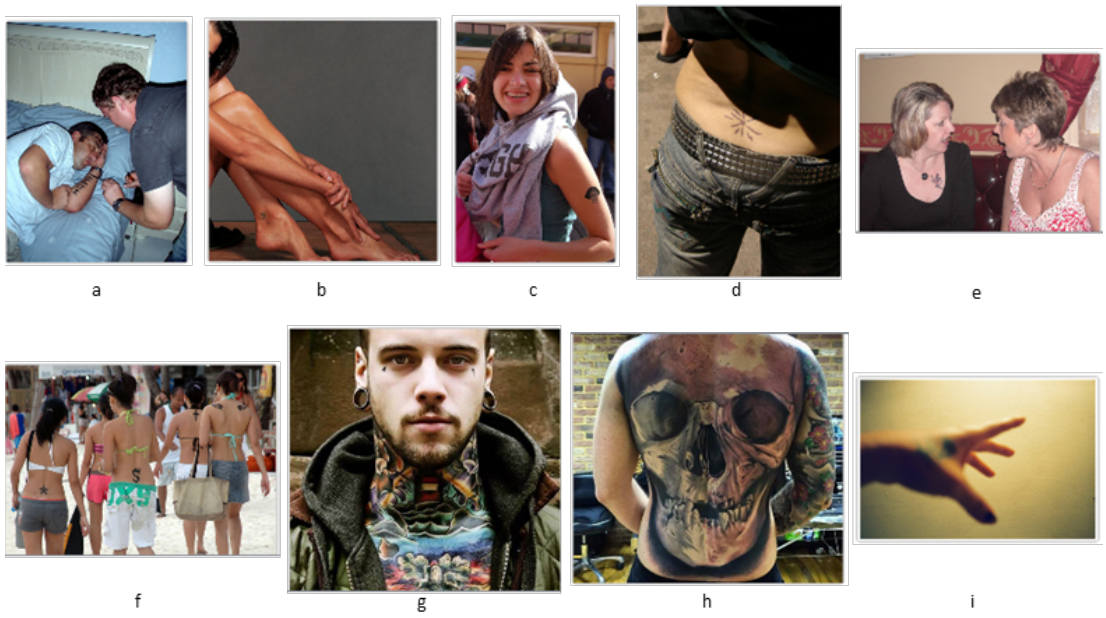

Figure 7. Examples of tattoos wrongly classified as non-tattoos

ACKNOWLEDGMENT

This study was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under research grant 311785/2019-0 to H.S. Lopes, and Fundação Araucária for the grant PRONEX-42/2018. Authors thanks NVIDIA for the donation of the Titan-Xp board used in this work.

REFERENCES

[1] H. A. Perlin and H. S. Lopes, "Extracting human attributes using a convolutional neural network approach," *Pattern Recognition Letters*, vol. 68, no. 2, pp. 250–259, 2015.

[2] Q. Xu, S. Ghosh, X. Xu, Y. Huang, and A. W. K. Kong, "Tattoo detection based on CNN and remarks on the NIST database," in *Proceedings of the International Conference on Biometrics (ICB)*, pp. 1–7, Piscataway, NJ, USA: IEEE Press, 2016.

[3] M. Ngan and P. Grother, "Tattoo recognition technology - challenge (Tatt-C): an open tattoo database for developing tattoo recognition research," in *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, pp. 1–6, 2015.

[4] M. Ngan, G. Quinn, and P. Grother, "Tattoo recognition technology - challenge (Tatt-C) - outcomes and recommendations," Technical Report NISTIR 8078, Gaithersburg, MD, USA, 2016.

[5] X. Di, , and V. M. Patel, "Deep tattoo recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 119–126, Piscataway, NJ, USA: IEEE Press, 2016.

[6] X. Di and V. M. Patel, "Deep learning for tattoo recognition," in *Deep Learning for Biometrics* (B. Bhanu and A. Kumar, eds.), Advances in Computer Vision and Pattern Recognition, pp. 241–256, Cham: Springer, 2017.

[7] X. Xu and A. W. K. Kong, "A geometric-based tattoo retrieval system," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 3019–3024, Piscataway, NJ, USA: IEEE Press, 2016.

[8] Z. H. Sun, J. Baumes, P. Tunison, M. Turek, and A. Hoogs, "Tattoo detection and localization using region-based deep learning," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3055–3060, Piscataway, NJ, USA: IEEE Press, 2016.

[9] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, pp. 2333–2348, 2019.

[10] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[11] M. Romero, M. Gutoski, L. T. Hattori, M. Ribeiro, and H. S. Lopes, "A study of the influence of data complexity and similarity on soft biometrics classification performance in a transfer learning scenario," *Learning and Nonlinear Models*, vol. 18, no. 2, pp. 56–65, 2020.

[12] M. Gutoski, M. Ribeiro, L. T. Hattori, M. Romero, A. E. Lazzaretti, and H. S. Lopes, "A comparative study of transfer learning approaches for video anomaly detection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 5, p. 2152003, 2021.

[13] M. Romero, M. Gutoski, L. T. Hattori, M. Ribeiro, and H. S. Lopes, "Soft biometrics classification in videos using transfer learning and bidirectional long short-term memory networks," *Learning and Nonlinear Models*, vol. 18, no. 1, pp. 47–59, 2020.

[14] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv preprint*, vol. 1602.07360v4, 2016.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," *arXiv preprint*, vol. 1512.00567v3, 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of 3rd International Conference on Learning Representations* (Y. Bengio and Y. LeCun, eds.), pp. 1–14, 2015.

[18] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.