Large Scale Data Collection of Tattoo-Based Biometric Data from Social-Media Websites

Michael Martin, Jeremy Dawson, and Thirimachos Bourlai
Lane Department of Computer Science and Electrical Engineering, West Virginia University
Morgantown, WV, 26505 USA
mmarti40@mix.wvu.edu, Jeremy.Dawson@mail.wvu.edu, ThBourlai@mail.wvu.edu

***Abstract*—The use of tattoos as a soft biometric is increasing in popularity among law enforcement communities. There is great need for large scale, publicly available tattoo datasets that can be used to standardize efforts to develop tattoo-based biometric systems. In this work, we introduce a large tattoo dataset (WVU-MediaTatt) collected from a social-media website. Additionally, we provide the source links to the images so that anyone can re-generate this dataset. Our WVU-MediaTatt database contains tattoo sample images from over 1,000 subjects, with two tattoo image samples per subject. To the best of our knowledge, this dataset is significantly bigger than any current released publicly available tattoo dataset, including the recently released NIST Tatt-C dataset. The use of social media in deep learning, data mining, and biometrics has traditionally been a controversial issue in terms of data security and protection of privacy. In this work, we first conduct a full discussion on the issues associated with data collection from social media sources for the use of biometric system development, and provide a framework for data collection. In this study, within the process of creating a new large scale tattoo dataset, we consider the issues and make attempts protect the subject's privacy and information, while ensuring that subjects remain in control of their data in this study and the use of the data adheres to the guidelines proposed by the Heath Care Compliance Association (HCCA) and the U.S. Department of Health & Human Services.**

***Index Terms*—Tattoo, Soft Biometrics, Social Media, WVU-MediaTatt**

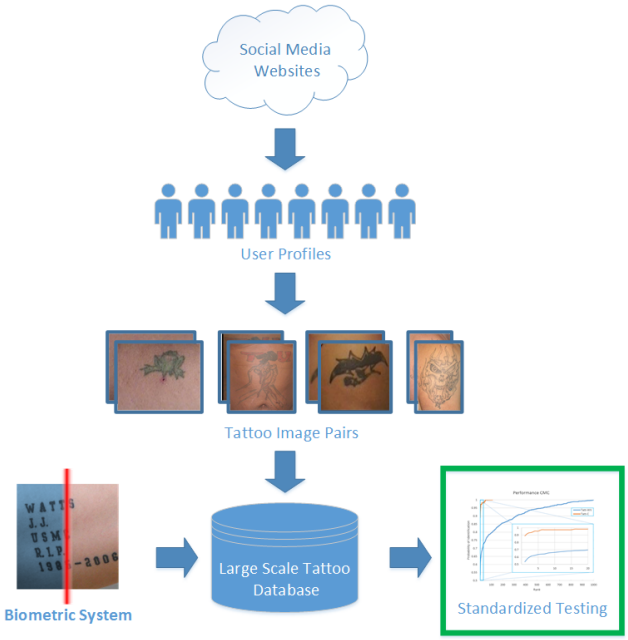

Fig. 1. **Public Test Standardization through Social Media Data Collection** Large scale data collections from social media sources allow for standardized testing of tattoo-based biometric systems on a scale not previously possible.

I. Introduction

Tattoos have shown to be an alternative to traditional biometric traits in solving some of the challenges associated with biometric identification. Tattoos are often classified as a soft biometric trait, as they contain less distinctiveness than traditional biometric traits (i.e. face, iris and fingerprint). Despite this categorization, several research studies have shown that they are useful for human identification in various law enforcement and security related applications [1]. One of the issues in tattoo-based human identification is the lack of publicly available tattoo datasets, which presents a great challenge in the development of such impactful and operational human recognition systems.

The use of online data in research applications is a commonly discussed topic and has shown to be of great benefit to a number of research applications. The Health Care Compliance Association (HCCA) defines a use of Internet Research as: "Research studying information that is already available on or via the Internet without direct interaction with human subjects (harvesting, mining profiling, scraping - observation or recording of otherwise-existing data sets, char rooms interactions, blogs, social media postings, etc.)" [12]. Furthermore, the collection of various tattoo-based image dataset from online sources has been reported to be very useful for the development of tattoo-based biometric systems [7][11][10]. However, most of the websites specializing in tattoo images are often intended for use as art galleries and, therefore, are unlikely to contain multiple images of the same tattoo. In order for tattoos to be used in a biometric system multiple images of the same tattoo are needed to form a gallery set of enrolled individuals and a probe set, which are query images that are matched against the gallery set in either verification or identification scenarios. One solution to the lack of sufficiently large tattoo datasets is the use of image transformations to generate simulated data by adding variance in illumination, blur, color, orientation, or change the size of the original limited size datasets, and thus, attempt to create suitable gallery and probe sets [2]. This method is effective at generating large tattoo datasets with multiple samples per

TABLE I
TATTOO DATA SOURCES

	Data Source	Type	Image Number	Multiple Samples	Availability	Previous Works
Private	Michigan State Police Dept.	Police Database	63,592	1,900 Duplicates	Not Public	[1] [2] [3]
	Thai Criminal Records Division	Police Database	444	177 Duplicates	Not Public	[4]
	Tatt-C / Tatt-E	NIST Database	4,332	109 Duplicates	Available with Approval	[5]
	WVU-Tatt	Privately Collected	940	79 Duplicates	Not Public	[6]
Public	Web-DP	Unknown	4,323	None	Public Website	[2]
	www.gangink.com	Police Database	256	None	Public Website	[7] [8] [9]
	www.eviltattoo.com	Art Gallery	9,631	None	Public Website	[10] [5]
	www.tattoodesign.com	Art Gallery	2,157	None	Public Website	[11] [10]
	www.checkoutmyink.com	Social Media	410,185	At Least 1,000 Duplicates	Public Website	Proposed

subject, although, the effects on performance of using multiple samples generated by image transformations need to be further explored to ensure their usefulness.

Conducting data collections (or partnerships with organizations with access to means of data collection) is a commonly used as an alternative to using online sources of data. One notable example of this is the use of tattoo images collected by the Michigan State Police Department [13][1][3]. Some of the advantages of conducting such data collections include the ability to strictly control the collection environment (i.e. lighting, background, distance, etc.), the involvement of the subjects (whether cooperative or non-cooperative), and the means of collection (using high end cameras or other data collection sensors). However, such studies can be very costly to conduct and often yield low number of test subjects unless the study is conducted over a long period of time (increasing the associated costs of collection). This is further amplified in tattoo-based biometric data collection due to ineligible portions of the population not having tattoos. According to a recent study conducted in 2013 by the Pew Research Center, only 14% of the U.S. population have at least one tattoo, thus making them eligible for the study.

The use of social media websites as a means of data collection could greatly increase the likelihood of finding multiple images of the same tattoo. In recent years, their use in biometric systems and data mining has become a controversial topic [14][15]. Information collected from social media is often used to identify an individual, extract or infer user preferences, or perform market analysis using data mining or pattern recognition techniques. In this paper, our discussion will be focused on the usage of tattoo images collected from social media sources for the development and evaluation of large scale, soft biometric recognition algorithms.

More specifically, in this work, we describe an alternative tattoo data collection process (shown in Fig. 1) involving the generation of tattoo datasets via a data collection toolkit (script) that connects to user-defined social-media websites. Using this process we were able to collect and generate a large scale tattoo dataset (WVU-MediaTatt) that far exceeds the size of other publicly available tattoo datasets suitable for use in the development and testing of tattoo-based biometric systems. Furthermore, given the ethical issues often involved with the use of social media data in biometric systems, special care is given to ensure individuals retain privacy and control of all information used in the creation of our WVU-MediaTatt dataset.

A. Related Works

Previous work focused on tattoo-related biometric systems have used data from a variety of sources with varying amounts of changes in background, illumination, and image quality. Several works [3][1][13] used a dataset provided by the Michigan State Police Department that, unfortunately, is not publicly available. This database has been used in biometric related works involving retrieval (tattoo recognition experiments), tattoo detection, segmentation.

Several non-social media based websites have been used for data collection as well. In some of these works [8][7][9] the website *www.gangink.com* was used. This website contains tattoo organized to specific gangs from near the Chicago area and could potentially be of use for content-related tattoo classification, however, its use in recognition based systems will likely be limited due to the lack of multiple images of the same tattoo. Other commonly used websites *www.tattoodesign.com* and *www.eviltattoo.com* are also unlikely to contain multiple images of the same tattoo, as they are intended to be used as art galleries for tattoo artists. Some of the online and privately-collected tattoo data sources used in the public works are shown in Table I.

The use of tattoos, traditionally a soft biometric trait, in recognition systems is rising as a viable alternative to more traditional biometric traits. Most of the previous work has focused on the task of tattoo image retrieval (also commonly referred to as identification). In a series of works by Jain et al. [11][3][13], a retrieval system called Tattoo-ID was proposed. This system used a matching technique that was based on local image features (i.e. SIFT). Another problem that is commonly addressed for tattoo related biometrics is segmentation. Segmentation can be defined as the isolation of image regions of interest from image background regions. One segmentation technique proposed by Kim et al. [16] involves the use of edge detection and morphological operators to extract the contours of the tattoo. This work has been recently expanded with new proposed methods of segmentation [5]. Using skin color intensity information, the segmentation region is then redefined to help eliminate background information. Allen et al. [7] proposed another segmentation method combining bottom-up and top-down cues. The work focused on first

performing skin detection to help localize tattoo regions, which can then be extracted using clustering, edge detection, or other target localization techniques. This method was tested with 256 tattoos collected from *www.gangink.com*. Another method for tattoo segmentation was proposed by Helfin et al. [17], where the authors proposed a combined face and tattoo recognition system where skin features (i.e. mole, scar, blemish, etc.) were also included. One of our previous works has focused on the use of the body location of the tattoo, fused with tattoo image matching to provide faster retrieval of tattoo [18].

The use of tattoo-to-sketch matching has also been explored in [9], and more recently, in [19]. Previous efforts in [9] were dependent on the use of local image features, however, new techniques developed by Huffman et al.) of performing image registration (tattoo to sketch alignment) combined with edge matching techniques have shown to be efficient in the 2015 National Institute of Standards and Technology (NIST) Tattoo Recognition Technology Challenge (Tatt-C).

II. Methodology

In this section, we will discuss how data collection is conducted and how we organize data into a usable state for identification-based biometric experiments. A full discussion is given on measures that we have taken to ensure privacy for the subjects included in the database that follows standards and principles proposed by the HCCA and HHS.

A. Data Collection

Firstly, to create a database of sufficient test size, a large number of images must be downloaded. In order to isolate groups of images that are likely to contain multiple image samples of the same tattoo, the data is organized by 'user profile'. Conveniently, the website *www.checkoutmyink.com* contains a members' section in which every user profile of the websites is listed. Currently the website contains over 170,000 users. A program is created to read the HTML webpage of the member index to search through the member list and download the public images available from each public profile. The pseudo code of this algorithm is shown in Algorithm 1.

With this algorithm, we were able to automatically collect over 117,000 tattoo images from over 20,000 profiles. However, not all of the collected images are suitable for use in a biometric database. Some of the images collected are not of a tattoo, but rather tattoo related equipment or art sketches. For this reason, much organization must be done to create a usable database. The process by which we manually organize our dataset from the automatically collected images will be discussed in the next section.

Algorithm 1: Data Collection Algorithm

```
Determine Number of Pages in Members Index (N) for i = 1 to N do
    Read HTML Page of Users
    Extract List of Profile Names with Length M
    for j = 1 to M do
        Read HTML Profile Page
        Extract List of Image Names with Length K
        for j = 1 to K do
            Save Image from Profile
        end
    end
end
```

B. Data Organization and Anonymization

In this section we will discuss how our data was processed and organized into a usable biometric database. In order to create a usable database, we must be able to organize subjects with multiple images per tattoo out of the 117,000 collected images. Firstly, any profiles that only contained one image were removed since they did not have the potential to have multiple images of the same tattoo. Some of the profiles created on *www.checkoutmyink.com* were advertisements for tattoo parlors or tattoo related equipment that must also be removed from potential test data. These profiles were relatively easy to locate because they often contained many images compared to a typical profile. The typical profiles were considered candidates for the WVU-MediaTatt database and were manually searched. Any pair of images that were of the same tattoo were organized as a subject into the WVU-MediaTatt database. We did not constrain a profile to contain a single tattoo pair, however, special care was given to ensure that a tattoo did not appear in the images used in multiple subjects. Using these techniques, we were able to organize the WVU-MediaTatt database to contain more than 1,000 subjects with two images per subject (for a total of over 2,000 tattoo images). This far exceeds previously publicly available tattoo databases used for biometric systems.

To ensure the anonymity of the profiles used in the creation of the dataset, the public links to the images are saved, but all information is removed. Each pair of images is then assigned a de-identified subject ID for reference in the WVU-MediaTatt database. This anonymization process works to remove the identifiable "private information" of each subject from the dataset. The Code of Federal Regulations, under Common Rule states that private information is considered identifying if the identity of the subject is or may readily be ascertained by the investigator or associate with the information". As we only release public links to the images on user profiles, no identifiable private information is communicated [20]. These image links can be acquired by contacting the authors at ThBourlai@mail.wvu.edu.

C. Dataset Evaluation

To evaluate our proposed dataset, we performed several recognition experiments in comparison with the Tatt-C dataset developed by NIST. Recently, Tatt-C has been used to benchmark several proposed tattoo-based systems for identification, detection, and content grouping.

Previous efforts to perform tattoo-based human recognition have been focused on the use of local image features, such as the SIFT algorithm. In accordance with these methods, our WVU-MediaTatt dataset and the Tatt-C are tested using the

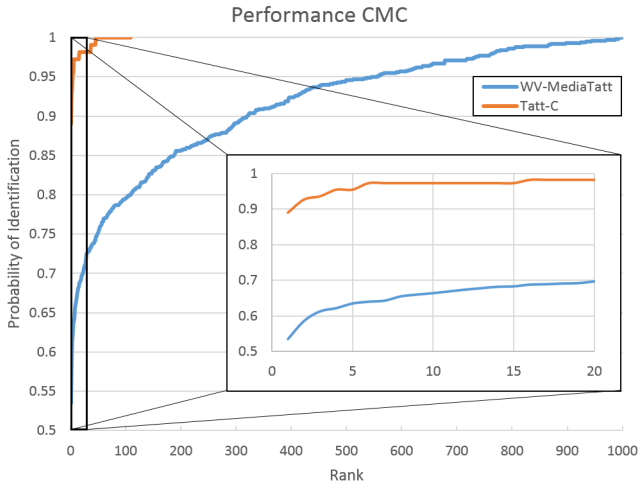

Fig. 2. **Performance Evaluation** We demonstrate the performance of our WVU-MediaTatt database (consisting of 1,000 tattoo images pairs) when used in recognition based experiments using local image features (i.e. SIFT). For comparison we evaluate Tatt-C database (consisting of 109 tattoo images pairs). Our newly proposed dataset poses a greater tattoo-based human recognition challenge, while Tatt-C can achieve relatively high recognition accuracy with simple matching with local image features.

OpenCV implementation of the SIFT algorithm. The CLAHE algorithm is used to perform photometric normalization prior to feature extraction. The L2 (Squared Euclidean Distance) is used as the distance metric between local image feature descriptors. The results of these experiments are shown in Fig. 2.

These results indicate that, while recognition methods using simple local image features may be sufficient to achieve highly accurate results (89.0% Rank 1), in instances with comprehensive and challenging datasets, more advanced matching techniques (i.e. deep learning based) may be needed.

III. Conclusion

In this paper we have present a new, publicly available tattoo database (WVU-MediaTatt) that can be used for identification-based biometric experiments. This WVU-MediaTatt database has been constructed from images collected from social media websites and contains only public links to these images on their original websites. By releasing the public links to these images only, we assist in protecting the confidentiality and identity of the subjects by ensure they retain full control of their image availability. Our database greatly exceeds the size of other publicly available tattoo databases suitable for identification (where two images of the same tattoo must be present), most notably NIST's newly proposed Tatt-C database. This database should greatly assist the field of tattoo-based biometrics by providing both a standardization in testing data and a large database in which to construct tattoo-based biometric identification systems.

A. Future Work

In future work, the techniques presented in this work could be expanded to gather data from any social media websites for a variety of biometric modalities. Currently the need to manually organize data and locate multiple images of the same tattoo from the subjects is a bottle neck in dataset creation. One solution to overcome this challenge would be to incorporate automated matching methods to locate subjects (profiles from the social media website) that are likely to contain multiple images of the same tattoo and organize them into a set that needs only to be reviewed by a user. Using these techniques, large datasets could be created with limited user input which is not possible on the same scale with manual data collection.

References

[1] J.-E. Lee, R. Jin, A. Jain, and W. Tong, "Image Retrieval in Forensics: Tattoo Image Database Application," *MultiMedia, IEEE*, vol. 19, no. 1, pp. 40–49, Jan 2012.

[2] J.-E. Lee, A. Jain, and R. Jin, "Scars, Marks and Tattoos (SMT): Soft Biometric for Suspect and Victim Identification," in *Biometrics Symposium, BSYM '08*, Sept 2008, pp. 1–8.

[3] A. Jain, J.-E. Lee, R. Jin, and N. Gregg, "Content-Based Image Retrieval: An Application to Tattoo Images," in *ICIP, 16th IEEE International Conference on*, Nov 2009, pp. 2745–2748.

[4] P. Duangphasuk and W. Kurutach, "Tattoo Skin Detection and Segmentation using Image Negative Method," in *ISCIT, 13th International Symposium on*, Sept 2013, pp. 354–359.

[5] J. Kim, H. Li, J. Ribera, and E. J. Delp, "Automatic and Manual Tattoo Localization," in *Technologies for Homeland Security, 2016 IEEE Conference on*.

[6] X. Xu, M. Martin, and T. Bourlai, "Automatic tattoo image registration system," in *2016 IEEE/ACM International Conference on ASONAM*, Aug 2016, pp. 1238–1243.

[7] J. D. Allen, N. Zhao, J. Yuan, and X. Liu, "Unsupervised Tattoo Segmentation Combining Bottom-up and Top-down Cues," in *SPIE DSS*, 2011, pp. 80 630L–80 630L.

[8] S. Acton and A. Rossi, "Matching and Retrieval of Tattoo Images: Active Contour CBIR and Glocal Image Features," in *SSIAI 2008. IEEE Southwest Symposium on*, March 2008, pp. 21–24.

[9] H. Han and A. Jain, "Tattoo Based Identification: Sketch to Image Matching," in *Biometrics, International Conference on*, June 2013, pp. 1–8.

[10] D. Manger, "Large-Scale Tattoo Image Retrieval," in *Computer and Robot Vision (CRV), Ninth Conference on*, May 2012, pp. 454–459.

[11] A. Jain, J.-E. Lee, and R. Jin, "Tattoo-ID: Automatic Tattoo Image Retrieval for Suspect and Victim Identification," in *Advances in Multimedia Information Processing PCM*, 2007, pp. 256–265.

[12] "Considerations and Recommendations Concerning Internet Research and Human Subjects Research Regulations, with Revisions," in *Secretary's Advisory Committee on Human Research Protections (SACHRP)*. Health Care Compliance Association (HCCA), 2013.

[13] A. Jain, R. Jin, and J.-E. Lee, "Tattoo Image Matching and Retrieval," *IEEE Computer*, vol. 45, no. 5, pp. 93–96, May 2012.

[14] B. Dana and C. Kate, "Critical Questions for Big Data," *Information, Communication and Society*, vol. 15, p. 5, 2012.

[15] J. M. Kleinberg, "Challenges in mining social network data: processes, privacy, and paradoxes," in *Proceedings of the 13th ACM SIGKDD*. ACM, 2007, pp. 4–5.

[16] J. Kim, A. Parra, H. Li, and E. J. Delp, "Efficient graph-cut tattoo segmentation," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 94 100H–94 100H.

[17] B. Heflin, W. Scheirer, and T. Boult, "Detecting and Classifying Scars, Marks, and Tattoos Found in the Wild," in *BTAS, IEEE Fifth International Conference on*, Sept 2012, pp. 31–38.

[18] M. Martin, X. Xu, and T. Bourlai, "A multimedia application for location-based semantic retrieval of tattoos," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–8.

[19] L. Huffman and J. McDonald, "Mixed Media Tattoo Image Matching Using Transformed Edge Alignment," in *Technologies for Homeland Security, 2016 IEEE Conference on*.

[20] "Code of Federal Regulations," in *Title 45: Public Welfare, Part 46 Protection of Human Subjects*. Department of Health and Human Services, 2009.