Looking for a Needle in a Haystack: How Can Vision-Language Understanding Help to Identify Privacy-Threatening Images on the Web

Sergej Schultenkämper
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
sergej.schultenkaemper@hsbi.de

Frederik S. Bäumer
Bielefeld University of Applied Sciences and Arts
Bielefeld, Germany
frederik.baeumer@hsbi.de

***Abstract*—Threats to user privacy in Web 2.0 are manifold. They can arise, for example, from texts, geoinformation, images, videos or combinations of these. In order to warn users of possible threats, it is necessary to find as much relevant information as possible. However, finding and aggregating relevant, user-specific information across web platforms, such as social networks, is challenging – not only because of the overwhelming amount of data but also due to the data quality and the great number of possible variants. In this paper, we investigate whether vision-language understanding techniques can be used to identify relevant image data and reliably extract sensitive information from these images. Our findings show that these methods are suitable for the pre-selection of relevant images, yet there are weaknesses in the extraction of information.**

***Index Terms*—Computer Vision; Privacy; Social Networks.**

I. Introduction

Users leave active and passive footprints through nearly every activity on the Web [1]. This includes quite obvious information, such as images, texts and videos that are knowingly uploaded by the users, as well as information that is passed on without the user's intervention, such as the IP addresses of the end devices or the user agent string. Furthermore, inherent information *hidden* in texts and images that are unknowingly published is difficult for users to keep track of. In the past, this has been demonstrated several times in an impressive and media-effective manner, such as by the automatic identification of vacation announcements and extraction of hidden Global Positioning System (GPS) image data on Twitter, which could for example be used to scout vacant properties for burglaries [2] or to reveal the running routes of soldiers on secret army bases, whose publication on sports portals revealed the exact location of the military installations [3]. It turns out that even small amounts of information can be dangerous in combination with other information [4].

In this paper, we focus on images with human attributes that are shared on the Web by users on different platforms and due to of different motivations – some of these images are meant to highlight a tweet, others are vacation or profile photos, while some are simply memes or photos of animals. From this fact comes the first challenge: Every day, millions of images are uploaded that pose no risk to users' privacy. Finding relevant images with human attributes in this huge, ever-growing and highly variable dataset is a complex task. Since we want to relate all the knowledge we get from an image to each other in order to extract reliable information, most classical image classification and segmentation methods fall short (e.g., limited domain, no extraction of class instances). We need an efficient, technical approach that enables sequential information extraction from images. For example, it is not sufficient to determine that a person has eyes and an eye color; rather, the specific eye color must also be reliably extracted (see Figure 1).

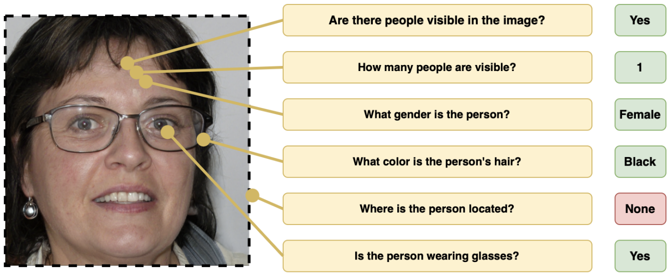

Are there people visible in the image?
Yes
How many people are visible?
1
What gender is the person?
Female
What color is the person's hair?
Black
Where is the person located?
None
Is the person wearing glasses?
Yes

Fig. 1. Attribute extraction via VQA

For this reason, we explore vision-language understanding techniques that can help to extract information from images. In particular, we are interested in techniques that allow Visual Question Answering (VQA), as this allows sequential questions and validation (in-depth questions, control questions). This approach is not to be considered in isolation, but can also be complemented by image classification approaches.

All of these considerations are taking place as part of the Authority-Dependent Risk Identification and Analysis in online Networks (ADRIAN) research project, which is dedicated to exploring and developing machine-learning-based methods for detecting potential threats to individuals based on online datasets and Digital Twins (DT). For this purpose, we discuss related work in Section II and describe the research concept and results of our privacy VQA approach in Section III. Finally, we discuss our findings in Section IV and draw our conclusions in Section V.

ISBN: 978-1-68558-069-8

II. Related Work

Here, we discuss the notion of DTs (see Section II-A) in the context of cyber threats and present related privacy research and image datasets (see Section II-B). Furthermore, we give an insight into Vision Language Models (see Section II-C).

A. Digital Twins in the context of cyber threats

The term DT is ambiguous and is used in a variety of areas in research and practice. It can be found in mechanical engineering, medicine, and computer science [5]. Developments in the field of Artificial Intelligence (AI) have given the term a wider usage. More generally, "DTs can be defined as (physical and/or virtual) machines or computer-based models that are simulating, emulating, mirroring, or 'twinning' the life of a physical entity, which may be an object, a process, a human, or a human-related feature" [5].

In the ADRIAN research project, we understand the term to mean the digital representation of a real person instantiated by information available on the Web [3]. In this context, the DT can never reflect the entire complexity of a real person, but reproduces features that, alone or in combination with other attributes, can pose a threat to the real person. In this way, the DT makes it possible to model the vulnerability of a person and make it measurable. The modeling of DTs is based on freely available standards of the semantic web, such as Schema.org and Friend Of A Friend (FOAF). This allows to connect and extend DTs. At the same time, the sheer number of possible sources of information, the quality of the data and a multitude of contradictory data make modeling challenging. However, studies show that a large amount of relevant information is knowingly and, to a large extent, unknowingly revealed by users themselves [4], [6]. It is precisely this fact that knowingly and unknowingly shared information on the Web can be merged and thus pose a threat to users which the ADRIAN research project aims to highlight [3].

B. Privacy research and image datasets

According to DataReportal [7], the average number of social media accounts per Internet user worldwide was 7.5 in 2022. The various Online Social Networks (OSNs) use mechanisms to protect the privacy of users. For user-generated content, such as user profiles (e.g., on Facebook), or geo-information (e.g., on Twitter), there are settings that can help protect this data. With regard to images, there are so far barely any options for protecting private visual information [8].

That said, DeHart et al. [9] processed Twitter data by analyzing texts and images in privacy context. Their study examines how users perceive privacy, how often privacy violations occur and what threats exist on Twitter. As for image analysis, the images were classified into three risk categories: "*severe*", "*moderate*" and "*no risk*". As a finding, images in the high risk category were found to contain primarily license plates, job offers and car keys. Moderate risk images are mainly images of job references, school information and promotion letters. The study confirms that depending on age, users are differently concerned about explicit websites, financial theft, identity theft and stalking. It also confirms that female and male participants are differently concerned about burglaries, explicit websites and identity theft.

Work already exists here that aims to help users preserve their own privacy. For example, Orekondy et al. [8] proposed a so-called Visual Privacy Advisor. This tool aims to assist users in enforcing their privacy preferences and preventing the disclosure of private information. They first create a dataset by annotating 68 personal information in images based on the EU Data Protection Directive 95/46/EC [10] and the US Privacy Act of 1974. Next, they conduct a user study to understand the privacy preferences of different users with respect to these attributes. They publish the Visual Privacy (VISPR) Dataset, which contains 22,167 images with a total of 115,742 labels. Finally, they extract visual features using CaffeNet [11] and GoogleNet [12], and train a linear Support Vector Machine (SVM) model. A final comparison between human and machine prediction of privacy risks on images shows an improvement of their model over human estimation.

In later work, Orekondy et al. [13] selected a subset of images of their VISPR Dataset for pixel-level annotation. This time, they focus on attributes that can be used for redaction so that the image is still useful. Reduction of a large building, such as a church, can make the image unusable. They propose the Visual Redactions Dataset with 8,473 images annotated with 47,600 instances for 24 attributes. The attributes are devided into the categories textual, visual and multimodal are then annotated. They also apply Optical Character Recognition (OCR) from Google Cloud Vision API to locate the text-based attributes. Furthermore, they apply Named Entity Recognition (NER) to recognize entity classes from the texts. As for visual attributes they apply models, such as Fully Convolutional Instance-Aware Semantic Segmentation Method (FCIS) [14] and OpenALPR to localize objects, such as faces, persons and license plates. Multimodal attributes are a combination of visual and textual information. Due to the limited amount of training examples and the large region of these attributes, they treat this as a classification problem. As a result, they propose a first model for automatic redaction of different private information.

Another system is presented by Spyromitros-Xioufis et al. [15]. This system performs privacy-aware classification of images. They created a dataset called YourAlert by asking users to provide privacy annotations for photos of their personal collections. The authors applied Latent Dirichlet Allocation (LDA) [16] to their corpus to identify the themes within annotations. In total, there were six topics related to privacy: "*Children*", "*Drinking*", "*Eroticism*", "*Relatives*", "*Vacation*" and "*Wedding*". They make the dataset publicly available with a total of 1,511 images, covering 444 private and 1,067 public images. Finally the VGG-16 model is applied to extract features, then they compute a modified version of the semfeat descriptor. The trained semi-personalized models lead to performance improvements over a generic model trained on a random subset of the PicAlert dataset.

Another relevant dataset is VizWiz-Priv [17]. The dataset consists of images taken by people who are blind to better understand the disclosure of their data. This dataset is used to develop algorithms that can decide first whether an image contains private information, and second whether a question about an image requires information about the private content of the image. A total of 8,862 regions including private content were tagged in the 5,537 images. When annotating the images, a distinction was made between private objects and objects that usually show private text. Images that show private objects consist of five categories, while images that contain private text consist of 14 categories.

C. Vision language models

In recent years, several vision language models, such as the Vision Transformer (ViT) [18], Contrastive Language-Image Pre-training (CLIP) [19], and Bootstrapping Language-Image Pre-training (BLIP) [20] have been published for multimodal deep learning. These models can be used to address various challenges in Computer Vision and Natural Language Processing. ViT is a type of neural network architecture designed specifically for image classification tasks [18]. It is based on the transformer architecture used in Natural Language Processing (NLP) models and uses self-attention mechanisms to process the image pixels in a parallel manner, allowing it to learn a rich representation of the relationships between different regions of the image [18]. ViTs have shown promising results in a variety of image classification tasks and have also been applied to other computer vision tasks, such as object detection and segmentation.

CLIP is a deep learning model for cross-modal representation learning. It learns a representation between natural language text and visual input (e.g., images) by comparing the similarity of the different image-text pairs [19]. The model has been trained on a dataset of 400 million image-text pairs collected from publicly available sources on the Internet [19]. The goal of CLIP is to create a representation that can be used for a variety of tasks, such as image captioning, VQA and text-to-image synthesis. CLIP is pre-trained on large amounts of text-image data and then fine-tuned on smaller task-specific datasets. This pre-training step helps the model learn a robust representation of the relationship between text and image, which can lead to improved performance on downstream tasks. CLIP consists of two encoders, a text encoder and an image encoder. The text encoder takes in a natural language text and produces a high-dimensional representation of the text. The text representation is generated by passing the text through a pre-trained language model. In CLIP, the text encoder is initialized with the pre-trained Bidirectional Encoder Representations from Transformers (BERT) weights [21]. The image encoder takes in an image and produces a high-dimensional representation of the image. The image representation is generated by passing the image through a pre-trained Convolutional Neural Network (CNN). Here CLIP uses a ViT or ResNet, depending on the task. The contrastive loss is used to train the encoders to generate similar representations for semantically related image-text pairs and dissimilar representations for semantically unrelated image-text pairs.

The authors of BLIP propose a new method to process noisy web data by bootstrapping the captions. It is called Captioning and Filtering (CapFilt) and improves the quality of the training data. Furthermore, they propose a multi-modal Mixture of Encoder-Decoders (MED), a multi-task model that can operate in one of three functionalities: unimodal encoder, image-grounded text encoder, and image-grounded text decoder [20]. The unimodel encoder for text and image is trained with an Image-Text Contrastive (ITC) loss. This functionality is the same as for the CLIP model pre-training. The image-grounded text encoder uses additional cross-attention layers to describe the interactions between image and speech and is trained with an Image-Text Matching (ITM) loss to distinguish between positive and negative image-text pairs [20]. Image-grounded text decoders replace bidirectional self-attention layers with causal self-attention layers and use the same cross-attention layers and feed-forward networks as encoders. For those given images, the decoder is trained with a Language Modeling (LM) loss to generate labels [20].

III. VQA Approach for Attribute Extraction

Our privacy VQA approach (see Figure 2) is straightforward and consists of three main steps: (1) data prepartion, (2) application of VQA, and (3) evaluation. In (1), we use the VISPR dataset [13], which consists of 67 different labels/privacy features. For our purposes, we select a subset of labels because not all labels are suitable for VQA processing, e.g., textual information, such as the full name or place of birth. We are primarily interested in directly visible personal attributes. In addition, we want to evaluate how different types of documents can be identified using VQA. The selected list of the attributes and documents is presented in Table I. The age group, gender, eye color, hair color and skin color have a large number of images with an average of 1,719 images. Followed by the five documents consisting of a national identification card, credit card, passport, driver's license and student ID, there are much fewer images with an average of 109. Further attributes of a person, such as tattoos, nudity and physical disabilities, are the least covered, with an average of 86 images.

In (2), after identifying and selecting the data, we perform an analysis of the different prompts using the BLIP model [22]. According to Jin et al. [20], prompts significantly affect zero-shot performance. We test different prompts, such as "*in the image*", "*in the photo*" and "*in the picture*". These different words have similar meanings, but it turns out that different word choices lead to different results. Next, we manually annotate all the selected attributes with labelstudio. Finally, we apply VQA to first detect images that contain a person. Here it is also important to determine how many persons are present in the image. The selected attributes can only be reliably extracted from images containing only one person. To do this, we use the following prompts: (I) "*Are there people in the picture?*", (II) "*How many people are in the picture?*" and (III) "*Is the face of the person visible?*"

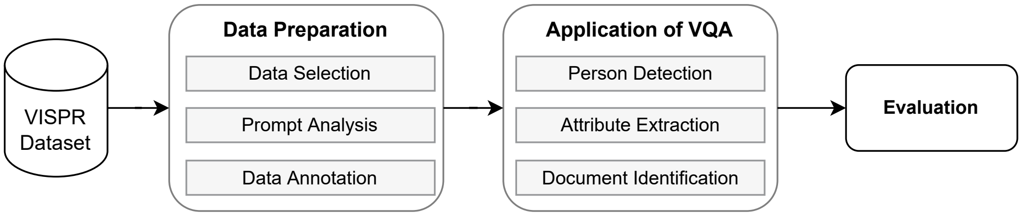


Fig. 2. Approach to VQA privacy analysis

TABLE I
SELECTED VISPR DATASET ATTRIBUTES

Attribute Id	Description	# of Images
a1_age_approx	Age Group	1711
a4_gender	Gender	1863
a5_eye_color	Eye Color	1348
a6_hair_color	Hair Color	1759
a11_tattoo	Tattoo	45
a12_semi_nudity	Semi-nudity	247
a13_full_nudity	Full nudity	11
a17_color	Skin Color	1914
a29_ausweis	National Identification	47
a30_credit_card	Credit Card	97
a31_passport	Passport	263
a32_drivers_license	Driver's License	70
a33_student_id	Student ID	70
a39_disability_physical	Physical Disability	41

TABLE II
QUESTIONS AND BLIP ANSWERS CANDIDATES

Label	Prompt & Answer Candidates
a1_age_approx	How old is the person? [*child, adult, elderly*]
a4_gender	What is the gender of the person? [*male, female*]
a5_eye_color	Which color are the eyes of the person? [*blue, green, gray, brown*]
a6_hair_color	Which color is the hair of the person? [*black, blond, brown, gray, red*]
a11_tattoo	Does the person have a tattoo? [*yes, no*]
a12_semi_nudity	Is the person partially nude? [*yes, no*]
a13_full_nudity	Is the person fully nude? [*yes, no*]
a17_color	What is the skin color of the person? [*black, brown, white*]
a29_ausweis, a30_credit_card, a31_passport, a32_drivers_license, a33_student_id	Which document is in this picture? [*national identification card*, *credit card*, *passport*, *driver's licence*, *student ID*]
a39_disability_physical	Does the person have a disability? [*yes, no*]

Furthermore, the focus on the extraction of personal attributes and the identification of documents. We use the prompts with the corresponding answer candidates shown in Table II. Most of the answer candidates are straightforward. For age, we have chosen to use child (up to about 16 years of age), adult (up to about 55 years of age) and elderly (55 years of age and over). We tested several alternative answer candidates and realized early on that if too many predictions are differentiated, such as middle-aged and old-aged adult, the annotation becomes very difficult because age can be very subjective. For the documents, we grouped all images and used only one prompt and all documents available in the dataset as answer candidates. To analyze attributes of a person in the context of yes/no answers, we added to each category the same number of images that did not belong to that label. To do this, we used images from the VISPR dataset with "*a0_safe*" labels , which indicate images that do not belong to any of the existing labels. Here, of course, it is equally important to see how well the model performs on images that are unrelated to the prompt. The final step is to evaluate the VQA performance.

To evaluate the privacy VQA performance of BLIP in (3), we used the Precision, Recall and F_1-score. In Table III, the results for the person attributs are shown.

For person detection, we took a random sample of 1,000 images. The detection of multiple persons is the most reliable with an F_1-score of 0.9757. After that, the images without persons are best recognised with an F_1-score of 0.9660. The model performs least well in detecting one person with an F_1-score of 0.9021. For personal attributes, the best results are obtained when classifying gender, age and skin color with F_1-scores between 0.9824 and 0.9346. This is followed by hair color and eye color with F_1-scores of 0.8865 and 0.8392. Furthermore, the detection of attributes, such as tattoos, semi-nudity, nudity and physical disability produces the worst results on average. Here, nudity detection achieves an F_1-score of 0.9565, but only for a very few images. This is followed by tattoo and semi-nudity with F_1-scores of 0.8222 and 0.8009 respectively. Physical disability only achieves an F_1-score of 0.7439. The results for the documents are shown in Table IV.

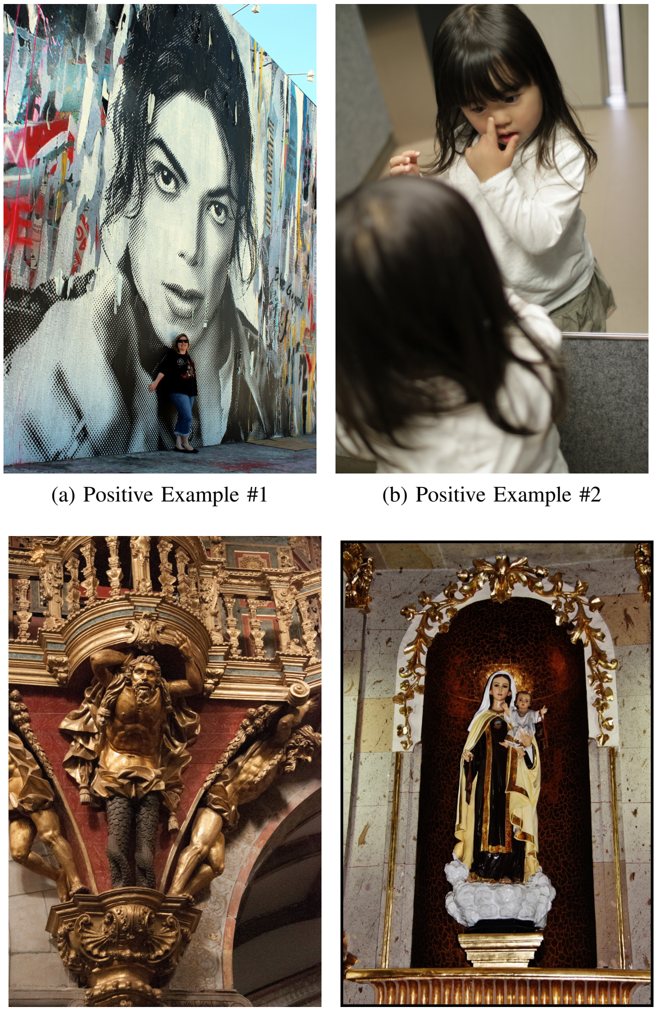
For document detection, we took the 547 available images of documents. For document identification, the best results were obtained for the passport, credit card and student ID with a F_1-score of 0.8529, 0.8468 and 0.6788, respectively. The driver's licence and national identification card yielded very poor results, with F_1-scores of 0.4268 and 0.3011, respectively.

TABLE III
PERSON ATTRIBUTE RESULTS

	Precision	Recall	F_1-score	Support
Person Detection				
No person	0.9977	0.9363	0.9660	455
1 person	0.8269	0.9923	0.9021	130
> 1 person	0.9730	0.9783	0.9757	369
Accuracy	–	–	**0.9602**	–
Age				
Adult	0.9853	0.9313	0.9575	1295
Child	0.9607	0.9293	0.9448	184
Elderly	0.6818	0.9626	0.7982	187
Accuracy	–	–	**0.9346**	–
Gender				
Female	0.9865	0.9787	0.9826	894
Male	0.9784	0.9862	0.9823	872
Accuracy	–	–	**0.9824**	–
Eye Color				
Blue	0.7415	0.7958	0.7677	191
Brown	0.8876	0.8791	0.8833	422
Gray	0.0000	0.0000	0.0000	1
Green	0.8000	0.2857	0.4211	14
Accuracy	–	–	**0.8392**	–
Hair Color				
Black	0.9749	0.9637	0.9692	523
Blond	1.0000	0.3457	0.5138	188
Brown	0.8416	0.9825	0.9066	687
Gray	0.8870	0.8160	0.8500	125
Red	0.6667	0.9630	0.7879	54
Accuracy	–	–	**0.8865**	–
Skin Color				
Black	0.8554	0.8256	0.8402	86
Brown	0.8015	0.7956	0.7985	137
White	0.9835	0.9859	0.9847	1635
Accuracy	–	–	**0.9643**	–
Tattoo				
no	0.8974	0.7447	0.8140	47
yes	0.7647	0.9070	0.8298	43
Accuracy	–	–	**0.8222**	–
Semi-Nudity				
no	0.8065	0.9375	0.8671	320
yes	0.7778	0.4930	0.6034	142
Accuracy	–	–	**0.8009**	–
Full Nudity				
no	0.9091	1.0000	0.9524	12
yes	1.0000	0.9167	0.9565	10
Accuracy	–	–	**0.9542**	–
Disability Physical				
no	0.6852	0.9024	0.7789	41
yes	0.8571	0.5854	0.6957	41
Accuracy	–	–	**0.7439**	–

TABLE IV
DOCUMENT RESULTS

	Precision	Recall	F_1-score	Support
Documents				
Credit Card	0.8557	0.8384	0.8468	99
Driver's License	0.5000	0.3723	0.4268	94
Nat. Ident. Card	0.2979	0.3043	0.3011	46
Passport	0.7719	0.9531	0.8529	213
Student ID	0.8000	0.8595	0.6788	95
Accuracy	–	–	**0.7148**	–



(a) Positive Example #1

(b) Positive Example #2

(c) Negative Example #1

(d) Negative Example #2

Fig. 3. Positive and negative examples

IV. DISCUSSION

All in all, the results show that our naive approach already leads to useful results, which can accelerate and improve the selection of relevant images. In particular, the important step of person detection has yielded good results. In the following, we discuss positive and negative examples (see Figure 3, a–d). As can be noted, there are some positive detections where it could be difficult for an AI model to identify the exact number of people that are present in the image. Examples are Figure 3 (a), which shows a woman standing in front of a large mural of Michael Jackson and Figure 3 (b), in which a little girl is standing in front of a mirror. In both cases, the image was classified as "*1 person*". As for the negative examples, there are many images of statues or emblems, that, for example, were classified as images with one (see Figure 3 , c) or more persons (see Figure 3 , d). While this can be considered as a *not completely wrong* classification, further experiments are necessary to find out how well real persons can be distinguished from statues, for example.

For the personal attributes, all cases achieved very good and usable results. It should be noted here, that the attributes "*age*" and "*hair color*" are very difficult to annotate. For "*age*", for example, it is very difficult to distinguish between an older adult and an elderly person without further knowledge. For "*eye color*", the annotators had to skip almost half of the images, despite the zoom function and high resolution of the images, because it was not possible to reliably determine the person's eye color. For the attributes with yes/no answers, "*nudity*" gave very good results and "*tattoos*" gave decent results. Both of these attributes are fairly easy to annotate. In case of "*semi-nudity*", it is difficult to determine where semi-nudity starts and where it ends. For example, according to the VISPR annotations a man with a naked torso is semi-nudity, the applied BLIP model mostly did not detect these cases.

For document identification task, "*passport*" and "*credit card*" are well detected as they do not differ much between countries. "*Driver's licences*" and "*national identification cards*" were very poorly identified by the model. Here, a detailed observation reveals a high variance in the representation of these documents across countries. We are currently working on an approach that currently only takes German documents into account in order to be able to develop country-specific approaches in further work, if necessary. However, we assume that in these cases a fine tuning of the models is necessary.

V. Conclusion

In this paper, we aimed to investigate whether an exemplary VQA method can help to preselect relevant images from a given dataset and extract certain human attributes. This could be an important pre-processing step in our research project ADRIAN, which aims to extract relevant attributes for different OSN users and initializes a DT.

As we were able to show, the BLIP model in its original form, i.e., without further fine tuning, can already demonstrate a very good detection rate for the number of people in an image and also shine in the recognition of human attributes. However, in terms of documents, the model is only suitable for identifying specific documents, such as credit cards and fails in detecting country-specific types of documents.

In future work, based on these initial results, we will focus on developing more advanced models on our annotated datasets in order to improve the results even further.

Acknowledgment

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, "Tracing Cross Border Web Tracking," in *Proceedings of the Internet Measurement Conference 2018*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 329–342.

[2] M. B. Flinn, C. J. Teodorski, and K. L. Paullet, "Raising Awareness: An examination of embedded GPS data in images posted to the social networking site twitter," *Issues in Information Systems*, vol. 11, no. 1, pp. 432–438, 2010.

[3] F. S. Bäumer, S. Denisov, Y. Su Lee, and M. Geierhos, "Towards Authority-Dependent Risk Identification and Analysis in Online Networks," in *Proceedings of the IST-190 Research Symposium (RSY) on AI, ML and BD for Hybrid Military Operations (AI4HMO)*, A. Halimi and E. Ayday, Eds., 10 2021.

[4] F. S. Bäumer, N. Grote, J. Kersting, and M. Geierhos, "Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews," in *International Conference on Information and Software Technologies*. Springer, 2017, pp. 77–89.

[5] B. R. Barricelli, E. Casiraghi, and D. Fogli, "A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications," *IEEE Access*, vol. 7, pp. 167 653–167 671, 2019.

[6] F. S. Bäumer, J. Kersting, M. Orlikowski, and M. Geierhos, "Towards a Multi-Stage Approach to Detect Privacy Breaches in Physician Reviews." in *SEMANTICS Posters&Demos*, 2018.

[7] Data Reportal, January 2022, https://datareportal.com/reports/digital-2022-global-overview-report, retrieved 05/25/23.

[8] T. Orekondy, B. Schiele, and M. Fritz, "Towards a Visual Privacy Advisor: Understanding and predicting privacy risks in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3686–3695.

[9] J. DeHart, M. Stell, and C. Grant, "Social Media and the Scourge of Visual Privacy," *Information*, vol. 11, no. 2, 2020.

[10] E. Directive, "95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the EC, L 281*, vol. 38, pp. 31–50, nov 1995.

[11] Y. Jia *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 675–678.

[12] C. Szegedy *et al.*, "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[13] T. Orekondy, M. Fritz, and B. Schiele, "Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8466–8475.

[14] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2359–2367.

[15] E. Spyromitros-Xioufis, S. Papadopoulos, A. Popescu, and Y. Kompatsiaris, "Personalized privacy-aware image classification," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. New York, United States: Association for Computing Machinery, Inc, jun 2016, pp. 71–78.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, p. 993–1022, 2003.

[17] D. Gurari *et al.*, "VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 939–948.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[19] A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[20] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren, "A Good Prompt Is Worth Millions of Parameters? Low-resource Prompt-based Learning for Vision-Language Models," in *Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*. Dublin, Ireland: ACL, May 2022, pp. 2763–2775.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, Jun. 2019, pp. 4171–4186.

[22] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *International Conference on Machine Learning*, 2022.