Applied Face Recognition in the Humanities

Martin Bullin[1], Andreas Henrich[1]

[1]University of Bamberg, Media Informatics, An der Weberei 5, 96047 Bamberg, Germany

Abstract

Several research areas in the humanities and social sciences could potentially benefit from state-of-the-art machine learning technologies. In the field of computer vision, there are exemplary face detection (FD) and face recognition (FR) algorithms that could provide support on several levels. We discuss two application scenarios where the deployment of trained networks can be used to generate further information. We show how FD can be used to recognize scene types such as dialogue and speech on non-photographic images such as Emblematica. The second part shows another application scenario where FR can be used to combine or link images in research datasets with authority records by finding personalities from the Wikidata dataset.

Keywords

Face Recognition, Face Detection, Scene Classification, Personality Mapping, Personality Recognition

1. Introduction

Searching collections of historical images and documents usually places high demands on labeling the collections with appropriate metadata. In this paper, we present first results of experiments in which we use face detection and recognition methods to firstly support a classification of the depicted scenes based on the detected faces and gaze directions and secondly to assign authority records (more precisely Wikidata items) to the depicted persons with as high accuracy as possible.

The first approach employs face detection algorithms for scene classification. Given a dataset, potentially comprising all images extracted from a particular book or collection, the objective is to classify the scene type present in each image. In our specific scenario, we focused on scene types such as *Dialogue*, *Speech*, and *Other*. The second proposed approach aims to simplify the process of labeling images by using metadata from a reliable reference source like WikiData. This reference source contains portrait images connected to authority records that can be identified using face recognition. By utilizing these available resources, the goal is to make the unambiguous labeling of the images in a research dataset easier and more efficient.

The remainder of the paper is organized as follows: In section 2 we will address related work and give more information on the background of our work. The use of face detection techniques for scene type classification is assessed in section 3. Section 4 considers the use of

LWDA'23: Lernen, Wissen, Daten, Analysen. October 09–11, 2023, Marburg, Germany

martin.bullin@uni-bamberg.de (M. Bullin); andreas.henrich@uni-bamberg.de (A. Henrich)

https://www.uni-bamberg.de/minf/team/bullin/ (M. Bullin); https://www.uni-bamberg.de/minf/team/henrich/ (A. Henrich)

0000-0001-9498-3615 (M. Bullin); 0000-0002-5074-3254 (A. Henrich)

face recognition techniques for labeling unseen data based on a reference source like WikiData and section 5 concludes the paper and presents potential future work.

2. Related Work and Background

In the upcoming subsections, an exploration of computer vision will be undertaken, focusing on three crucial areas: face detection, face recognition, and scene classification. Face *detection* algorithms are designed to locate and identify human faces within images or video frames. Face *recognition* techniques enable the identification and verification of individuals based on their unique facial features. Lastly, *scene classification* methods, which will be further discussed in subsection 2.2, aim at the categorization of images or video frames into various scene types.

2.1. Face Detection and Face Recognition

In the domain of face detection and its subfield, face recognition, a wide range of tools are available. Wang et al. conducted a comprehensive survey [1], wherein they evaluated 15 different methods. The results demonstrated that all of these methods achieved accuracies exceeding 99% on simple datasets, and at least the superior methods, attained high accuracies above 90% across all datasets. Due to the focus of our investigations on the application of state-of-the-art methods, the selection of tools was initially guided by their user-friendly nature, followed by their cutting-edge performance and significance within the research community.

A well utilized tool in the field is InsightFace[1], which encompasses various CNN based components such as RetinaFace for face detection and ArcFace for face recognition [2]. With over 15k stars on GitHub, this project has gained significant popularity and is available as a Python library. Its utilization, particularly for face detection and during the deployment phase, is relatively straightforward. However, the training process for the standard implementation presents greater complexity, as it requires the dataset to be in the MXNet binary format. Hence, in section 3 where face detection was necessary, InsightFace, particularly RetinaFace, was employed. Conversely, it could not be employed in section 4 that involved the learning of new faces for the MWW dataset which consists of old engravings with portraits. The tool employed in this application scenario was the face_recognition library[2]. Unlike InsightFace, this library offers ease of use for both the deployment and training aspects of face recognition tasks. In default mode it works without CNN models and thus can be run on CPU solely. It is developed based on the dlib C++ library and claims to achieve a remarkable accuracy of 99.38% on the *Labeled Faces in the Wild*[3] dataset. With 50k stars the corresponding repository enjoys even greater popularity than InsightFace. Figure 1 gives examples of the face detection quality for both libraries. It will be discussed in more detail in section 4.3.

[1] https://insightface.ai/ (all URLs accessed in July 2023)
[2] https://pypi.org/project/face-recognition/
[3] http://vis-www.cs.umass.edu/lfw/

(a) ID 16 (b) ID 1821 (c) ID 2811 (d) ID 5815 (e) Gates 1 (f) Gates 2 (g) Gates 3 (h) Gates 4

Figure 1: Examples of the MWW-Portrait Dataset and Bill Gates images with the detected faces generated by the two used face detection algorithms. (face_recognition above with blue markings, InsightFace below with red markings)

2.2. Scene Classification

In the field of "Scene Classification", it is noteworthy to mention that no publications specifically addressing this particular use-case were found. Most publications focus on annotated standard datasets like the *MIT Indoor 67* with predefined classes like *indoor* and *outdoor* with multiple subclasses like *store* or *home* [3, 4, 5, 6]. The majority of these publications and subsequent studies employ deep learning models to classify these standard scene classes [7]. However, only one publication by Wevers et al. [8] introduces classes that are specifically relevant to the *dialogue* and *speech* scenario. Further research could explore testing our approach against the performance of these existing models on their dataset, limited to the defined classes.

Another research field delves into the specific topic of *dialogue detection*, with a particular emphasis on this singular class. In the pioneering work of Kotti et al. [9], audio-assisted dialogue detection in movies was explored, marking one of the early instances of deploying neural networks in this context. The exploration of our proposed approach with video data could shed light on the suitability and efficacy within this specific research domain.

Another topic is described by Impett showing an approach of clustering the gestures found in art history [10]. In future the mentioned automatic human pose and gesture estimation could be compared to as well as combined with the proposed approach.

An approach that already bridges the domains of dialogue detection and classification with face recognition is presented in the work by Ito et al., which uses additionally sound [11]. The authors aim to recognize *smile* and *laughter* by combining speech processing and face recognition techniques. Similar to our proposed approach, they utilize facial landmarks generated from the face as part of their methodology. This prior work serves as a relevant reference in demonstrating the feasibility of integrating facial landmark analysis into the context of dialogue classification.

Our study aims to address the research gap in scene recognition, specifically focusing on the recognition of *dialogue* and *speech*. To achieve this, the study leverages state-of-the-art deep learning approaches in combination with a rule-based classification.

3. Face Detection Deployment for Scene Type Detection

Face detection has advanced significantly, leading to the development of various applications such as InsightFace-REST[4]. These applications encompass a range of functionalities, including face detection, face recognition, age estimation, and even specialized tasks like mask detection during the COVID-19 pandemic. The algorithms utilized in these applications often rely on facial landmarks and bounding boxes to perform their tasks effectively.

Facial landmarks, represented by red and green dots in the lower part of Figure 1, are key points on the face used in face detection and recognition algorithms. They typically include landmarks corresponding to the eyes, nose, and mouth regions. By leveraging the positions and relationships between these landmarks, algorithms can accurately detect and recognize faces.

In this study, the generated facial landmarks and bounding boxes were employed to infer the scene type depicted in the underlying image. Various scene types, as previously mentioned, were introduced for classification purposes. The scene type *Dialogue* specifically refers to an interaction where two individuals are engaged in a conversation, facing and looking at each other. Another scene type, *Speech*, is defined by the presence of a speaker making eye contact with at least one other person, with the number of people observing the speaker surpassing a customizable threshold. Images not classified as *Dialogue* or *Speech* encompassing scenes that do not fit the specific criteria are categorized as *Other*.

This study highlights how existing deep learning solutions can be leveraged to address problem scenarios that extend beyond their primary purposes. In this particular case, the focus was on utilizing a face detection algorithm to infer the scene types depicted in images. By exporting the results and generating a script, the study demonstrates how the different scene classes can be effectively described and captured.

In the forthcoming sections, the concept of the study will be elaborated upon in detail, outlining the proposed methodology. Subsequently, a brief introduction to the examined datasets will be presented. Finally, the qualitative evaluation conducted on the provided datasets will be depicted, highlighting the findings and outcomes of the study.

3.1. Concept

Acknowledging that the application scenario is hypothetical and emphasizing the exploration of possibilities in addressing new problem scenarios, it is important to note that the generated classes, namely *Dialogue* and *Speech*, are also constructed and leave space for individual interpretations and refinements. While a simple solution could involve counting the number of faces in an image and using a threshold to determine whether it is a *dialogue* or a *speech* (i.e., two faces indicating a dialogue and more than two faces indicating a speech), our approach considers additional features. We used the Oxford English Dictionary (OED) definitions as a basis for *Dialogue* and *Speech*. According to the OED, a *Dialogue* is defined as a "Conversation between two or more characters in a literary work; the words spoken by the actors in a play, film, etc."[5] In the context of this research, the focus was specifically on two-way conversations, where two individuals are present and maintain visual contact by looking at each other. On

[4] https://github.com/SthPhoenix/InsightFace-REST

[5] https://www.oed.com/search?searchType=dictionary&q=dialog&_searchBtn=Search

the other hand, *Speech* is defined as an "address or discourse of a more or less formal character delivered to an audience or assembly."[6] Our approach takes into consideration factors such as the potential eye contact between more than two people, the number of people looking at a single person, and the person in focus looking into the direction of most individuals present.

To determine the main auxiliary variables, namely gaze-direction, eye-contact, and looks-at, for the detected faces generated by the face detection network, the following process was implemented:

Gaze-direction: The gaze-direction for each face is classified as either *left*, *right*, or *not-clear*. The threshold for *not-clear* is set at the minimum, to start with a high recall, but keeping the possibility to move from a binary classification into *left* and *right* to a more precise one including *not-clear*. This determination is based on the relative position of the point between the two eyes with respect to the vertical bounding box borders. If the point is closer to one of the vertical bounding box borders the gaze-direction is assigned as the direction of the closer bounding box border.

Looks-at: The gaze-direction information is then used to check if a person is looking at another person. The looks-at variable is set to true if a person is located beside another person and the gaze-direction indicates that he or she is focusing on one of the keypoints of the other person. This calculation is based on the viewing angle, which is defined by the vector connecting the eyes and an angle of default 180° "around" this vector. The choice of a relatively high default value, exceeding the biological norm, was made due to the limitations of the 2D representation of images and the unconventional nature of the datasets.

Eye-contact: The eye-contact variable summarizes whether both person 1 and person 2 have the "looks-at" flag for each other, indicating mutual eye contact.

Using these collected pieces of information for all faces in the image, a categorization process was implemented using a simple Python script. This script facilitates the classification and analysis of the facial features and interactions within the image with a simple heuristics. To categorize the scenes into *Dialogue*, *Speech*, or *Other*, the following steps are undertaken:

Derivation of centrality from a Looks-at statistics: The list of looks-at pairs is split into all occurrences of individuals involved. Assume a scene where person a is looking at person b and at person c, and person c is also looking at person a. Then from the list of looks-at pairs $[ab, ac, ca]$, the list of all occurrences of individuals $[a, b, a, c, c, a]$ would be derived. This helps to identify the most prominent or central person in the image by assessing the frequency of their appearance in the next step.

Frequency counter: A frequency counter is created to determine the occurrence frequency of each person in the list containing all occurrences of individuals determined before. Referring to the example this would look like $[a{:}3, b{:}1, c{:}2]$. This provides insights into the relative prominence of individuals within the scene.

Threshold calculation: A threshold value is computed by multiplying the adjustable threshold value (0.7 in our example) by the number of individuals looking at someone minus one. This adjustment accounts for the fact that the speaker must look at someone. In the given example this would be $0.7 \cdot (3 - 1) = 1.4$. So at least 2 looks would have to go from or to the speaker.

After these preparation steps, the classification is performed by the following checks:

[6] https://www.oed.com/view/Entry/186128

Dialogue classification: The first check involves verifying if only two persons have mutual eye contact and only these two persons look at someone leading to the classification *Dialogue*.
Speech classification: The next check ensures that there is at least one eye-contact present and that the person with the highest number of looks exceeds the calculated threshold. Then the class *Speech* is applied.
Other classification: In all other cases the class *Other* is assigned.

3.2. Data

To address the lack of annotated datasets for the task at hand, three different datasets were considered for evaluation:

Emblematica: The Emblematica Online Collection[7] is a challenging dataset in this study. It consists of 1,388 facsimiles of emblem books from various libraries. Emblems within this collection typically comprise different components, such as a heading, one or more mottos, and a pictura, which will be classified by the approach.

MWW Portraits: The MWW Portraits collection[8] contains approximately 32,000 prints, including 6,000 duplicates, dating from the 16th to the mid-19th century. It encompasses various printmaking techniques and primarily features scholars, bourgeoisie, and portraits of theologians.

Jürg Straumann Artwork: The dataset comprises digitized artworks by Jürg Straumann[9].

These three datasets were selected because they represent non-photographic images related to the humanities and offer a diverse range of potential application scenarios due to their variations in content and characteristics.

3.3. Evaluation

Due to the absence of labeled datasets, a quantitative evaluation could only be performed on a subset of the results. Additionally, it is important to note that the interpretation of the results may vary, as there is no definitive ground truth. It is crucial to consider the specific application scenario when assessing the correctness of the results.

For instance, in Figure 2 (a), an image depicts two dialogues: two individuals in the background engaged in a conversation and two individuals in the foreground. Although the image was classified as a dialogue by the system, the prediction might be seen as incorrect. However, if the main objective of the application scenario is to identify images with dialogues, allowing for the presence of more than two individuals, the classification may also be considered appropriate.

In such cases, it may be worthwhile to adjust the decision process of the algorithm by implementing a more suitable approach. This could involve reducing the threshold for the face probability, which would increase the recall but potentially lower the precision. It is important

[7] https://emblematica.grainger.illinois.edu/
[8] https://vfr.mww-forschung.de/portraetsammlung
[9] https://juergstraumann.ch

Table 1
Resulting image classification counts of the approach for the three investigated datasets.

Class	Emblematica	MWW	Straumann
Dialogue	464	317	11
Speech	394	324	4
Other	948	1443	26

to strike a balance between maximizing recall and maintaining an acceptable level of precision, considering the potential trade-off of classifying non-face regions as faces.

The results presented in Table 1 indicate that the system classifies less than 50% of the images into one of the meaningful classes. This can be attributed to several factors.

Firstly, the incompleteness of the class set plays a significant role. Since the classes were generated for the purpose of this study and are not derived from a comprehensive or predefined set, there may be instances where the classes do not fully encompass the variations and nuances present in the images.

Secondly, the algorithm itself is designed to prioritize accuracy, aiming to provide reasonably precise classifications. This focus on precision may result in a higher number of images being classified as *Other* due to the conservative nature of the algorithm.

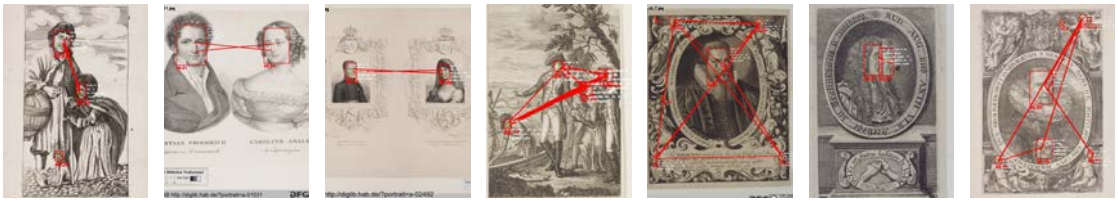
In the following exemplary qualitative assessments of the results for the three datasets will be given.

Emblematica When examining specific examples, such as those depicted in Figure 2, the results appear to align quite well with the defined classes. For instance, when analyzing the first 30 images in alphabetical order of the image names classified as dialogues, 27 of them accurately reflect the definition, resulting in a true positive rate of 90%.

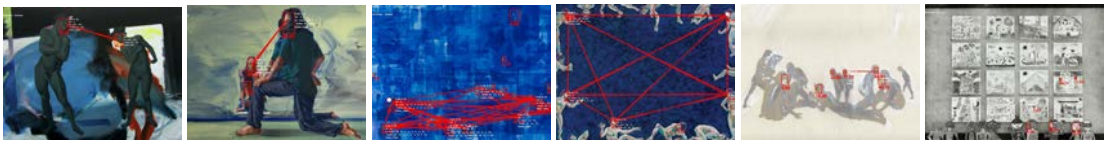
Applying the same evaluation method to the class *Speech*, the accuracy rate varies depending on the interpretation of *Speech*. If *Speech* is understood as the presence of a main actor with people looking at him or her, the accuracy rate is approximately 83% for the first 30 images. However, if *Speech* is interpreted according to the defined criteria, the accuracy rate may be lower at around 67%.

(a) Dialogue Unclear (b) Dialogue Unclear (c) Dialogue Correct (d) Dialogue Correct (e) Speech Correct

Figure 2: Exemplary results by the scene classification algorithm for both target classes out of the Emblematica dataset.



(a) Dialogue Correct (b) Dialogue Correct (c) Dialogue Unclear (d) Speech Correct (e) Speech False (f) Others Correct (g) Others Correct



(h) Dialogue Correct (i) Dialogue Correct (j) Speech Unclear (k) Speech False (l) Others Unclear (m) Others Correct

Figure 3: Exemplary results by the scene classification algorithm for all classes out of the MWW Portrait (a – g) and Straumann Artwork (h – m) Datasets.

Similarly, when considering the *Other* images, the results are comparable to those of the *Speech* class. Most of these images do not neatly fit into either of the two defined classes, and their classification may depend on the interpretation of *Speech*.

MWW The nature of portraits typically implies that only one person is prominently depicted, which aligns with the expectation that a majority of the images would be classified as *Other*. This is evident from the results presented in Table 1, where a larger proportion (69%) of the MWW Portraits dataset is classified as *Other* compared to the Emblematica (52%) dataset.

Interestingly, the results for *Dialogue* and *Speech* in the MWW Portraits dataset still show fitting classifications, albeit dependent on the specific definition used. To provide a comprehensive view, Figure 3 displays exemplary results for all classes. This allows observers to form their own opinion and interpretation based on their own understanding and definition of the classes.

Straumann In the case of the Straumann artwork dataset, the face detection algorithm performs good, but the detection probabilities are often relatively low. A stringent reduction of the dataset takes place, when the images are filtered by the face probability threshold. It is important to note that this reduction process was applied to all datasets for the purpose of comparison and to ensure consistent evaluation. Despite the reduction, the remaining images in the dialogue class as exemplarily depicted in Figure 3 demonstrate a good fit, indicating accurate classification. However, the results for the speech class are not as precise, which can be attributed to the unique nature of the artwork dataset. Interpreting scenes in artwork can be challenging, particularly for individuals without art-related expertise. Nevertheless, the correct classifications obtained for this specialized dataset are noteworthy, highlighting the potential for accurate scene recognition even in complex and nuanced scenarios.

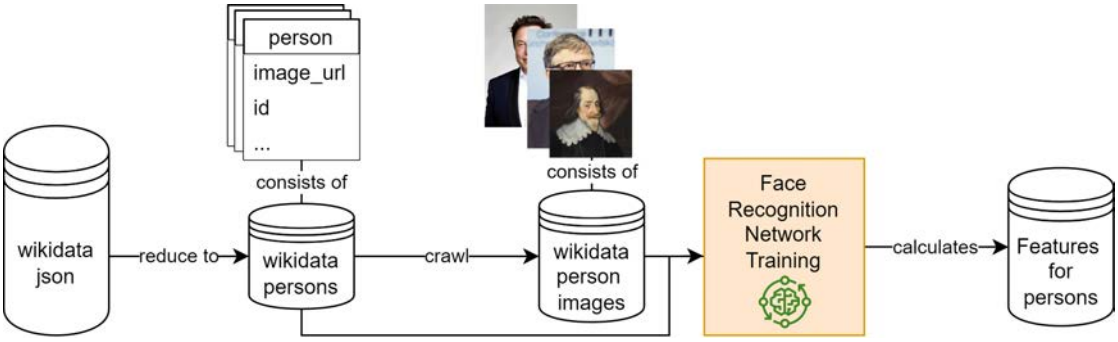


Figure 4: Training process for face recognition trained on Wikidata

4. Face Recognition Deployment for Labeling Unseen Data

This section focuses on the practical implementation of face recognition techniques for labeling unseen data and mapping datasets of persons with images that may have different naming conventions but represent the same individuals. The section is structured into three subsections: The concept section provides an overview of the underlying principles and methodologies. In the data section, the training dataset, namely Wikidata, is discussed. Lastly, the evaluation section outlines the metrics and methodologies used to assess the performance of the proposed approach.

4.1. Concept

Due to the very good results of the face detection algorithm used in section 3, the corresponding face recognition algorithm would have been the first choice for the subsequent section. The high effort to preprocess the images into the training format, lead us to the choice of the face recognition network developed by Adam Geitgey[10] as outlined in Section 2. It is superior in the context of ease of training and deployment. Based on the crawled person dataset explained later and shown on the left side of Figure 4, the feature representation for each image was calculated, as shown on the right side of the figure. This representation of persons is then used to identify the best-matching faces in the dataset when presented with query images, as can be seen in Figure 5. The comparison returns a list of calculated distances; the smaller the distance the more similar the images tend to be. In the example image the ice hockey player Kevin Gloor with the Wikidata id *Q1740152*[11] has the smallest distance to the search image of Chester Bennington.

4.2. Data

The Wikidata[12] platform serves as a valuable starting point for this project. As stated on their website, Wikidata is a free and open knowledge base that allows both humans and machines to read and edit its structured data.

[10]https://github.com/ageitgey/face_recognition
[11]https://www.wikidata.org/wiki/Q1740152
[12]https://www.wikidata.org/wiki/Wikidata:Main_Page

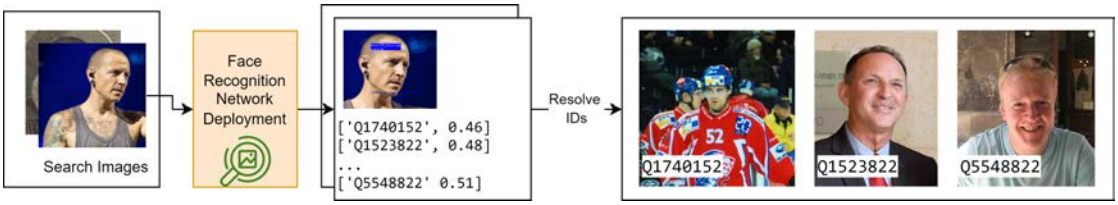


Figure 5: Recognition process for face recognition trained on Wikidata showing the results

Utilizing the whole Wikidata dump, a dataset comprising 1,109,006 human entities with 1,141,894 associated images was generated. The images for each person were crawled in a descending order based on the number of images available in Wikidata to increase the chance of hitting a matching image, if the person is included in the dataset.

The process began with the complete Wikidata JSON dump (the left database symbol in Figure 4). This compressed file of about 110 GiB was filtered to include only the ID and image URL of entities representing humans, identified by the *Q5* ID with the *P18* property denoting the presence of an image URL. Through the original Wikidata IDs in the resulting 80MiB file, we could expand the process by filtering to a specific time period by using the original Wikidata dump.

4.3. Evaluation

The analysis primarily focused on a qualitative assessment to evaluate the performance and limitations of the deployed network. This involved testing both, images gathered from the internet and non-photographic images sourced from the MWW portrait dataset[13] to pseudo-quantitatively examine different types of images.

Photo Detection: Bill Gates To assess the qualitative precision and recall of the face recognition architecture, a person from the crawled dataset with multiple images in the Wikidata dataset was selected for analysis. In this case, Bill Gates was chosen. The four images depicting him in Wikidata can be found in Figure 6 on the left side. To challenge the face recognition network, ten images with various differences to the ground truth images were selected as query images[14]. The differences included wearing sunglasses, without glasses, different mouth positions during speech, high contrast due to sunlight, a gray-scale image, and even an image of Bill Gates Sr. to check, if similarity of relatives can be detected as well.

The network successfully identified Bill Gates among the top 10 similar faces for five out of the ten query images. Additionally, for eight out of the ten images, the network detected Bill Gates within the top 100 similar faces. Only two images posed bigger challenges for the network: The image of Bill Gates Sr., which, considering the difficulty in recognizing the similar facial features due to their familial relationship, was expected. The second was the one where he was depicted with Melinda under high contrast conditions. Interestingly, Melinda herself was recognized with the lowest difference among all faces in that particular image.

[13] https://vfr.mww-forschung.de/portraetsammlung
[14] https://www.google.com/search?q=bill+gates&tbm=isch

(a) 4 Ground Truth Images (b) 10 Query Images

Figure 6: Ground truth images (from Wikidata) and query images of Bill Gates for the evaluation of the FR algorithm.

The image of Bill Gates with sunglasses and without glasses yielded similar results, with the FR network identifying faces containing different types of glasses or no glasses at all. The gray-scale image predominantly generated results containing other gray-scale images, while the image of Bill Gates Sr. resulted in visually similar depictions of older men. These outcomes align with common assumptions about the network's ability to recognize facial features and characteristics despite variations in accessories, image color, and familial resemblance. These results highlight the capabilities and limitations of the FR network in accurately identifying and matching faces.

Non photographic Dataset: MWW Portraits To evaluate the performance of the face recognition system on the MWW portrait dataset, the ten most prominent persons were identified (most frequent in the MWW portrait dataset), and the corresponding depicted images were mapped to their respective Wikidata IDs manually resulting in 8 personalities with over 450 test images for evaluation. Unfortunately, despite expanding the search space to the 500 closest matching faces when searching the Wikidata images, not a single portrait image was correctly classified. The observed performance issues of the chosen face recognition algorithm on the MWW Portraits, can also be seen in in Figure 1, where it is compared to the InsightFace approach, leading to the conclusion that alternative face recognition algorithms may offer better results in recognizing faces from non-photographic datasets. The complexity involved in generating training datasets using the original implementation, which led to the use of this FR algorithm, could be addressed using third-party implementations of InsightFace yielding an easier to train solution.

An alternative approach could involve exploring the possibility of retraining or replacing the relevant components of the face_recognition library. Further investigation into the dlib Library reveals that feature comparison is performed using a stored neural network. Because the documentation does not provide detailed insights into this process, the existence of numerous alternatives implies that exploring other options could easier yield improved results.

The results presented in this section suggest that the recognition of people in photos works well enough to support an annotation process. However, the library used was not convincing on drawings and engravings, for example. Unfortunately, the original InsightFace library used

in the section 3 could not be used in this scenario for the technical reasons mentioned above. However, InsightFace already shows in Figure 1 and also in the scenario considered in section 3 that there is clear potential with suitable libraries for non-photographic images.

5. Future Research

The investigated research scenarios presented in the previous sections were designed to demonstrate and prove the applicability of face detection and recognition methods in applications from the humanities. During the investigations it turned out that the potential usefulness of the methods can in addition be demonstrated by a scenario that emerged in discussions with researchers from the social sciences. It involves the analysis of images crawled from school websites. In this scenario, the first stage would involve checking the images for the presence of faces in order to filter for relevant images. Subsequently, face recognition could be employed to identify unique faces and obtain frequencies for age, gender, and potentially ethnicity with additional efforts. The images could then be classified into categories such as "class pictures", "portraits" or "events" using a similar approach to the one applied in section 3, providing further insights into the dataset and the way schools present themselves on the web.

This scenario underscores the importance of user-friendly tools for researchers in the humanities and social sciences. Furthermore, it emphasizes the need for continuous improvement and refinement of the methodologies employed.

In the following sections, we will outline some potential enhancements that could be implemented to improve the two methodologies.

5.1. Face Recognition for Data Labeling

One potential avenue for future research is to explore the use of alternative face recognition tools and techniques. Specifically, investigating networks that are trained on multiple images of the same person could potentially improve the usability and accuracy of the face recognition process. Additionally, incorporating methods developed under the InsightFace project, which have demonstrated better performance on non-photographic datasets, particularly for face detection, could further enhance the applicability of the face recognition component.

Furthermore, for a real-world application scenario involving the integration of metadata from Wikidata, implementing a filtering mechanism based on additional criteria such as the years in which the dataset's content was generated could be beneficial.

5.2. Scene Recognition

A potential improvement for scene recognition could be to leverage the generated ground truth, and perform a manual annotation process to specifically train a neural network. By using this ground truth data, it is probably possible to train a network that can achieve even better classification performance.

Additionally, there is the opportunity to optimize the approach by further refining the heuristic algorithm for determining the classes. Taking into account the relative sizes of faces in an image could provide additional contextual information for scene classification. Furthermore,

refining the algorithm's ability to accurately determine the gaze-directionof persons could contribute to more precise scene interpretations.

Overall, the topics explored in this study and the results offer valuable insights and inspire potential research questions. It is noteworthy that with a network capable of comparing an input image to over a million images, the detection of a person within the top-100 results is achieved in 8 out of 10 instances. This shows the potential for exciting application scenarios.

References

[1] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, Y. Guo, A Survey of Face Recognition, 2022. arXiv:2212.13038.

[2] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5202–5211. doi:10.1109/CVPR42600.2020.00525.

[3] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 413–420. doi:10.1109/CVPR.2009.5206537.

[4] J. Wu, H. Christensen, J. Rehg, Visual Place Categorization: Problem, Dataset, and Algorithm, in: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, 2009, pp. 4763–4770. doi:10.1109/IROS.2009.5354164.

[5] S. Lazebnik, C. Schmid, J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, 2006, pp. 2169–2178. doi:10.1109/CVPR.2006.68.

[6] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene recognition from abbey to zoo, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3485–3492. doi:10.1109/CVPR.2010.5539970.

[7] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen, L. Liu, Deep Learning for Scene Classification: A Survey, 2021. arXiv:2101.10531.

[8] M. Wevers, Scene Detection in De Boer Historical Photo Collection:, in: Proceedings of the 13th International Conference on Agents and Artificial Intelligence, SCITEPRESS - Science and Technology Publications, Vienna, Austria, 2021, pp. 601–610. doi:10.5220/0010288206010610.

[9] M. Kotti, E. Benetos, C. Kotropoulos, I. Pitas, A neural network approach to audio-assisted movie dialogue detection, Neurocomputing 71 (2007) 157–166. doi:10.1016/j.neucom.2007.08.006.

[10] L. Impett, Analyzing gesture in digital art history, in: The Routledge Companion to Digital Humanities and Art History, Routledge, 2020, pp. 386–407.

[11] A. Ito, X. Wang, M. Suzuki, S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video, in: 2005 International Conference on Cyberworlds (CW'05), 2005, pp. 8 pp.–444. doi:10.1109/CW.2005.82.