Recognizing Characters in Art History Using Deep Learning

Prathmesh Madhu
Pattern Recognition Lab,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
prathmesh.madhu@fau.de

Ronak Kosti
Pattern Recognition Lab,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
ronak.kosti@fau.de

Lara Mührenberg
Institute of Church History,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
lara.muehrenberg@fau.de

Peter Bell
Institute for Art History,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
peter.bell@fau.de

Andreas Maier
Pattern Recognition Lab,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
andreas.maier@fau.de

Vincent Christlein
Pattern Recognition Lab,
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
vincent.christlein@fau.de

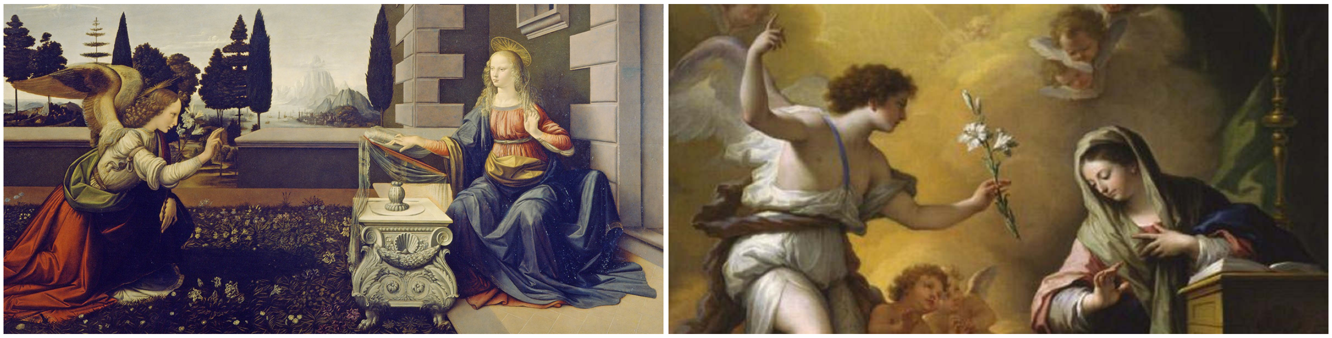


Figure 1: Art historical scene depicting the iconography called *Annunciation of the Lord* (left [10], right [32]). Mary and Gabriel are the main protagonists. We can clearly see the differences in the background, in the artistic style, in the foreground, in the objects, their properties, and the use of color.

ABSTRACT

In the field of Art History, images of artworks and their contexts are core to understanding the underlying semantic information. However, the highly complex and sophisticated representation of these artworks makes it difficult, even for the experts, to analyze the scene. From the computer vision perspective, the task of analyzing such artworks can be divided into sub-problems by taking a bottom-up approach. In this paper, we focus on the problem of recognizing the characters in Art History. From the iconography of *Annunciation of the Lord* (Figure 1), we consider the representation of the main protagonists, *Mary* and *Gabriel*, across different artworks and styles. We investigate and present the findings of training a character classifier on features extracted from their face images. The limitations of this method, and the inherent ambiguity in the representation of *Gabriel*, motivated us to consider their bodies (a bigger context) to analyze in order to recognize the characters. Convolutional Neural Networks (CNN) trained on the bodies of *Mary* and *Gabriel* are able to learn person related features and ultimately improve the performance of character recognition. We introduce a new technique that generates more data with similar styles, effectively creating data in the similar domain. We present experiments and analysis on three different models and show that the model trained on domain related data gives the best performance for recognizing character. Additionally, we analyze the localized image regions for the network predictions.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Image representations**; **Machine learning**; • **Applied computing → Arts and humanities**.

ACM Reference Format:
Prathmesh Madhu, Ronak Kosti, Lara Mührenberg, Peter Bell, Andreas Maier, and Vincent Christlein. 2019. Recognizing Characters in Art History Using Deep Learning. In *1st Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3347317.3357242

1 INTRODUCTION

The understanding of scenes in art images, specifically called as iconography, is a very demanding task due to the complexity of the scene. Interpreting a scene involves detection and recognition

SUMAC ’19, October 21, 2019, Nice, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6910-7/19/10…$15.00
https://doi.org/10.1145/3347317.3357242

of various present objects, their relationship to each other, their impact on the visual perception of the scene and their relevance for the respective task. In images and paintings from various artworks, the understanding of a scene becomes more complex even for an expert for the following reasons: *(1)*, the artworks are an artistic representation of real-life objects, people and scenes or inspired by these; and *(2)* the artistic style differs from one artist to another and it also may change from one painting to another, even for an artwork of a given theme [7–9]. For example, in Figure 2 we can see Mary (a - d) and Gabriel (e - h) represented by different artists and their respective styles. Interesting to note here is that all of the images are from the same iconography called *Annunciation of the Lord*, however their representation differs from one image to another. Sometimes the artworks are fragmented or have unique compositions. Artists quite often employed different means to convey a message. For example, they would use the body pose of the main protagonists (observe the differences in the body pose of Mary and Gabriel in Figure 1) or even gestures [4, 23].This makes the interpretation of an artwork more difficult, even for an expert. From computer vision perspective, this problem poses an interesting challenge for vision techniques to understand an artwork with some objectivity. Understanding art involves complex processes such as discovering and recognizing the background structure of the scene, the objects present and their significance within the scene, their relationship with the main object of focus, artistic style and the higher level semantics to deconstruct the meaning of the artwork. With recent advances in computer vision, human-level performance has been achieved in many tasks like detection, recognition and segmentation of objects on natural real world images [28]. However, understanding the underlying semantic knowledge still remains an open challenge [53].

Our work on recognizing character in artworks attempts to explore recognition features about the main subjects depicted in the given scene. This analysis is of great importance for art history. For example, Figure 2 shows examples of *Mary* and *Gabriel*. They are the main protagonists and hence the center of focus in the iconography called *Annunciation of the Lord*. Figure 2 (a – d) shows Mary in the female form, however, the representation of Gabriel (e – h) is not always clear. In humanities, opinions about Gabriel’s gender are divided, with some claiming that it is male while others arguing that it has no gender at all. This is because Angels are regarded as beings created by God, mediating between the heavenly and earthly spheres. The corporeal nature of angels is discussed in detail in the Judeo-Christian tradition [46]. Early Christianity attributed an etheric body to them [39], and in the High Middle Ages, the doctrine prevailed that angels had a purely spiritual body [25]. Unlike humans, angels do not have a sex or gender. In the biblical stories, the angels appear in human bodies (more precisely male), whose appearance they can assume. They are therefore also called “sons of God” or “sons of heaven”. The name of the Archangel Gabriel (Hebrew ”גַבְרִיאֵל“) contains the word stem ,גבר which means “man” or “strength” [33]. In the book of Daniel, his appearance is described as “like a man” (Dan 8:15). In art, angels are first shown in human male form, analogous to the biblical findings [20]. Their fiery nature (ether) can be made visible by the red colouring of their skin [39], their belonging to the heavenly sphere by bluish translucent flesh [20]. From the Middle Ages onward, the asexuality of the bodies of angels was sometimes more strongly depicted [24]. The angel Gabriel himself therefore has no sex or gender, but manifests himself in a human appearance, which can have male and female parts. These can also be made visible in art. Due to this ambiguity it is very difficult to classify Gabriel as purely male or female. In this work, we aim to solve a sub-problem of recognizing *Mary* and *Gabriel* within artworks. In order to recognize the characters of Mary and Gabriel, we introduce approaches that can accomplish this task within art images, regardless of the type of art, its style or the artist.

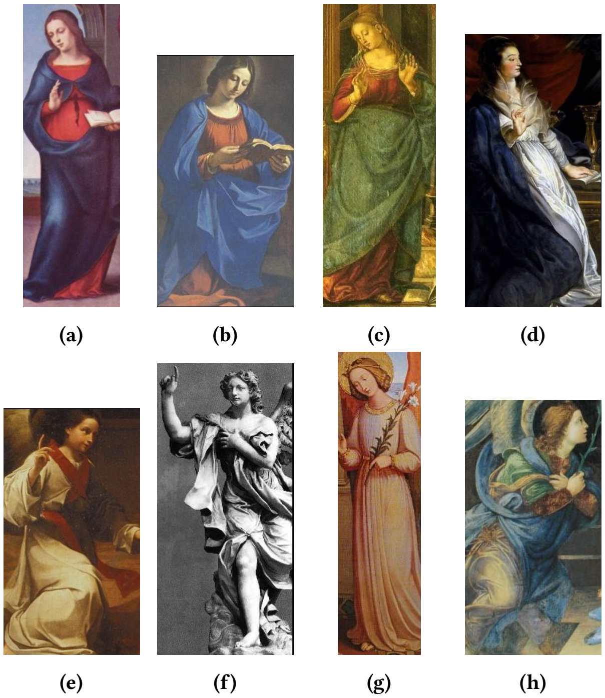


Figure 2: Depiction of Mary (a – d) and Gabriel (e – f) as the main protagonists in the *Annunciation of the Lord* iconography

In the last few years, several methods have been introduced for object recognition (including people) on real-world datasets in computer vision. For example, traditional methods using handcrafted features, such as filter-based feature extractors [31], which can be aggregated using Bag of (visual) Words [15]. Current state-of-art systems use CNN-based deep networks as object feature extractors [11, 29, 51]. However, the protagonists represented in the artworks have very different and unique characteristics compared to photographs (as shown in Figure 3). Hence, it becomes very important to use the domain knowledge present in the artworks for recognizing character.

The majority of the existing techniques in computer vision [13, 36, 37, 41] rely on facial features in order to recognize the person. Since the artworks are representations of the artist’s imagination and skill, it is difficult to recognize the person from their facial features alone. A case in point would be that of Gabriel. As seen in Figure 2(e - h), Gabriel’s face gives an impression of being a boy (e), man (f), a woman (g) or the perception is unclear (h). The use of only faces for recognizing character in artworks is therefore insufficient. In our proposed method, we use the entire body images of Mary and Gabriel to model the problem of recognizing character.

In addition, it also would be interesting to find out the semantic understanding of the trained model when it recognizes a particular character from an artwork. Usually, in artworks, the face is used in addition to other meta-information such as clothing or the neighboring contextual information to analyze the paintings. Figure 1 shows two images from the iconography of *Annunciation of the Lord*. We can see that there is an angel-like figure to the left with a pose pointing to another figure on the right of the image. The figure on the right is that of Mary. It is represented as a female form, while the left figure is that of the Archangel Gabriel, whose representation seems neutral without apparent gender reference. In order to achieve a higher semantic understanding of the scene depicted here, it is important for the vision model to recognize these characters.

In this paper we demonstrate that recognizing specific characters (Mary and Gabriel) is possible using deep models. Our contributions are as follows: *(1)* we show that the performance of traditional machine learning models, such as SVMs (Linear and RBF kernels), Logistic Regression and Random Forests decreases when they are trained on the whole body of the characters as opposed to only faces; *(2)* we introduce a novel technique as a way of simplifying the transfer of knowledge from one domain to another; *(3)* we show that this technique is beneficial for the model's performance and outperforms traditional machine learning algorithms.

The paper is organized as follows: In section 2, we discuss about the related work in person identification and transfer learning that could be useful for recognizing characters in art; In section 3, we introduce traditional as well as deep learning based methods adopted by us for recognizing characters and also furnish the details about the dataset preparation; Section 4 essentially details the experimental setup for all the methods and their corresponding quantitative and qualitative evaluations; and In section 5 we make a small discussion and conclusions about our adopted methods and their respective merits.

2 RELATED WORK

Person identification has been a core problem in computer vision. Since the advent of video surveillance technologies, cheap hardware for recording and storing videos and the latest research in object recognition techniques, it has become important to automatically identify a person in an image or a video to keep a track on the movements for security reasons. Many common methods use facial features for the recognition of a person. Usually, the character (or person) identification is divided into two parts: *(1)* feature extraction and *(2)* feature classification that recognizes the character's identity. In traditional computer vision techniques, features are hand-crafted, e.,g., LBP, HOG, SIFT or an ensemble of these local descriptors to encode the information present in the face [6, 42–44]. On the other hand, the feature classifier methods include neural networks [18], Adaboost [3] and Support Vector Machines (SVM) [34]. With the advent of the recent end-to-end deep learning algorithms, raw face pixel intensity values are given as input to the CNN-based classifiers for training and then these trained classifiers are used to assign a class to an unseen face image. CNNs use convolutional filters to learn the facial features by training millions of example images [14, 36]. These methods provide a unique advantage of training a network in an end-to-end fashion.

Table 1: Face Datasets

	IMDB-Wiki	LFW-Face	Adience	CelebA
Source	Wikipedia	Web	Flickr	Google
# Images	62,328	13,233	26,580	202,599
# Identities	20,284	5749	2,284	10,177
Reference	[41]	[21]	[13]	[30]

The implementation of deep models was particularly successful due to the availability of large datasets. Table 1 shows the details of available face datasets with their respective sources, number of images, and number of identities. The number of images typically needed to train a CNN from scratch is huge and in the range of millions [26]. However, the face datasets (Table 1) do not have such large amounts of data. This situation poses a dilemma: Since the deep networks need large amounts of data to train from scratch, the available face data is insufficient in comparison.

Transfer learning is a method where the knowledge learned in one domain can be carefully transferred to another, which can be beneficial [35, 47]. Typically, the new domain to be adapted for transfer learning must have a similar feature space and data distribution as the original domain. The current deep networks that are trained on person or face recognition have used transfer learning by using pre-trained networks to jump start the training [36, 40, 50]. These pre-trained networks provide a good starting point for training without requiring large amounts of data.

Recent works applying modern computer vision methods to the problems related to artworks claim that the images or paintings they work with differ substantially in semantics and representations from natural images [7–9]. For example, Figure 3 shows sample images from IMDB-Wiki & Adience datasets [13, 41] (row 1). In comparison to Figure 2, it is apparent that the artworks show a very different character representation. They could be inspired from real-life or often a manifestation of the artist's imagination.

Recent research on the semantic understanding of paintings in art, however, shows that it is possible to apply transfer learning of vision to artworks while training new models. For example, Garcia and Vogiatzis [16] considered the Text2Art challenge where they successfully retrieved the related artwork from the set of test images 45.5% (Top-10) of times. They also showed a technique to store visual and textual information of the same artwork in the same semantic space, thereby making the task of retrieval much easier. In another instance, Zhong et al. [50] showed that it is possible to model a deep feature network on a particular artwork collection using self-supervised learning for discovering near-duplicate patterns in larger collections.

Strezoski and Worring [45] also showed that using multi-task learning techniques with deep networks for feature learning gives

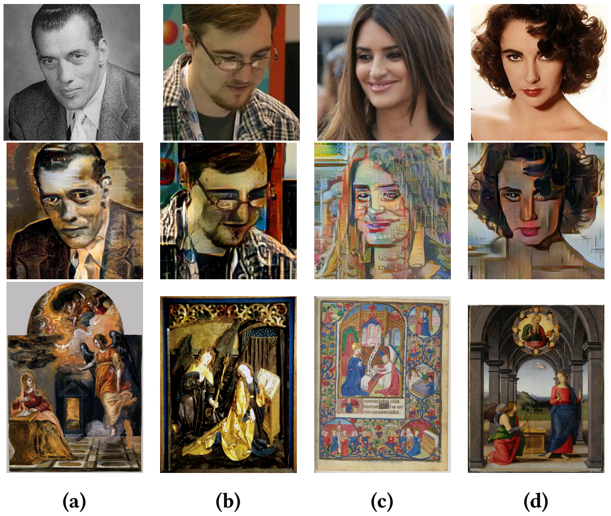
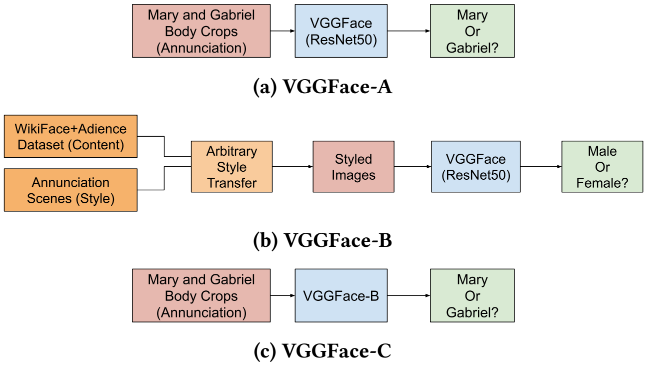


Figure 3: Face images taken from [13, 41] in row 1 with their style-transferred counterparts in row 2 and their corresponding style images in row 3. Columns (a) & (b) depict images of men whereas columns (c) & (d) depict women.


Mary and Gabriel Body Crops (Annunciation)
VGGFace (ResNet50)
Mary Or Gabriel?
WikiFace+Adience Dataset (Content)
Annunciation Scenes (Style)
Arbitrary Style Transfer
Styled Images
VGGFace (ResNet50)
Male Or Female?
Mary and Gabriel Body Crops (Annunciation)
VGGFace-B
Mary Or Gabriel?

(a) VGGFace-A

(b) VGGFace-B

(c) VGGFace-C

Figure 4: Pipelines for Recognizing Characters. The model definitions of VGGFace-* are in section 3.2

better performance over hand-crafted features. Clearly, it is possible to use techniques like transfer learning from state-of-art in computer vision to adapt for art related problems.

3 METHODOLOGY

In this section, we explain how we created our database of annunciation scenes and its related style-transferred dataset. Afterwards, the models used to train on this dataset are described.

3.1 Database Creation

Our dataset[1] consists 2787 images [1, 2, 17, 27, 48] of artworks from the iconography called *Annunciation of the Lord*, with focus on the main protagonists: *Mary* and *Gabriel*. The image data is from a corpus of medieval and early modern annunciations, acquired from public domain. We generated bounding box estimates for the bodies of both using a fast object detector called YOLO [38]. These were corrected manually by art history experts. The distribution of Mary and Gabriel is nearly balanced with 1172 and 1007 images, respectively. We also generated bounding boxes for their faces using an image annotation tool called VIA [12]. This database served as a source for all the experiments mentioned in section 3.2, except VGGFace-B.

Table 1 shows some currently available datasets in computer vision for identity recognition of real-life individuals. We chose to use IMDB-Wiki [13] and Adience [41] since the images are not only limited to faces, but also contain the upper body part. Combining both, we have around 20000 images belonging to each class, male and female, which we call the *content images*. Since there are fewer female samples, we have chosen a similar number of male samples. Figure 3 (row 1) shows sample images of two men and two women from the combined dataset. Our annunciation dataset has 2787 images which we call the *style images*. Figure 3 (row 3) shows four sample annunciation scenes serving as style images, from our database. Using a style transfer model, based on adaptive instance normalization, introduced by Huang and Belongie [22], we transferred the artistic style of *style images* to the *content images*. For each style image, we transfer its corresponding style to 16 (8 males and 8 females) randomly selected content images. In this way, we are able to use each individual style of the annunciation scenes, resulting in a similar distribution of styles in the style-transferred images. These images have a similar style to the annunciation scene, effectively making the distribution of our content data similar to that of our style data (annunciation scenes). Figure 3 (row 2) shows some samples of the style-transferred images. The styles of row 3 have been transferred to the corresponding images in row 1. Since the content images have categorical labels of *male* and *female*, we now have a similar styled, larger dataset, which can be used to learn the domain knowledge of the annunciation scenes. This database served as a source for training the VGGFace-B model (see below).

For the annunciation dataset, we used 200 test images, 100 each for Mary and Gabriel and the rest for training, making a split of approximately 90%/10% between train/test. For the styled dataset, we used a split of 75%/25% between train/test.

In the next subsection, we introduce methods used for training on these datasets.

3.2 Proposed Methods

In all our proposed methods, we use a ResNet50 model [19], pretrained on the VGGFace dataset [36]. This model was trained to identify people from their face images, so it's feature space is very well trained for recognizing characters.

For recognizing characters, we experiment with the following five methods:

(1) ML-Face: We take a pretrained ResNet50 [19], trained on object recognition and remove the top-most softmax layer. Thus, we use the activations of the penultimate layer (2048 dimensional vector) as our features. These are extracted for all the face images of Mary and Gabriel. Using these as input

[1] data acquired for non-commercial scientific research

Table 2: Performance metrics for traditional machine learning models trained on the features from *FACE* images of Mary and Gabriel.

Model Type	Pr	Re	F1	Acc.
Random Forests (200 est.)	0.75	0.75	0.75	0.75
Logistic Regression	0.70	0.68	0.68	0.68
SVM (Linear, C = 100)	0.78	0.78	0.78	0.78
SVM (RBF, C = 1000, γ = 0.01)	**0.80**	**0.79**	**0.79**	**0.79**

features, we train Random Forests, SVMs (Linear and RBF kernels) and Logistic Regression.

(2) ML-Body: The examples of Figure 2 show the differences in the perception of Mary and Gabriel. The face does not have all the information to perceive the differences between Mary and Gabriel. Hence, we propose to include more contextual information in the form of their bodies. For this experiment, we take the bodies of Mary and Gabriel and extract their ResNet50 features similar to ML-Face.

(3) VGGFace-A: For this experiment, we also take a pretrained ResNet50 [19] trained on the VGGFace dataset [36] and replace the top-most 1000-class softmax layer with a 2-class sigmoid layer. Since Gabriel's gender is ambiguous, we used sigmoid for final layer activations. The receptive fields of CNNs are able to learn the contextual and hierarchical information [52], we fine-tune this model on the body images of Mary and Gabriel. Figure 4 (a) shows a visual representation of the model.

(4) VGGFace-B: Once more, we take a pretrained Resnet50, similar to the previous methods. However, we directly train it on the styled image dataset. Instead of Mary and Gabriel, we take Female and Male as the corresponding labels since we expect the model to learn the styles rather than person related features. Figure 4 (b) shows a visual representation of the model.

(5) VGGFace-C: We take VGGFace-B from the above method and fine-tune it on the Mary and Gabriel bodies dataset. We expect that VGGFace-C will be able to learn specific features related to Mary and Gabriel because VGGFace-B has been trained on an annunciation style related dataset. Figure 4 (c) shows a visual representation of the model.

4 EVALUATION

In this section we explain the experiments conducted, their training steps, results and analysis for all the models introduced in subsection 3.2. All the results in the form of tables or otherwise have been calculated on the images of the test set, i. e., the models have not been trained on any image of the test set. The CNN models were trained by using standard hyperparameter values for learning rates, dropout rates, batch sizes and epochs. We augment the data on the fly with techniques like shear, shift, horizontal flip and rotation. These techniques ensure that the model is not biased towards specific poses or orientation of the characters.

Table 3: Performance metrics for different models trained on the features from *BODY* images of Mary and Gabriel. The model definitions of VGGFace-* are given in section 4.

Model Type	Pr	Re	F1	Acc.
Random Forests (200 est.)	0.59	0.56	0.54	0.59
Logistic Regression	0.68	0.68	0.68	0.69
SVM (Linear, C = 10)	0.68	0.68	0.68	0.68
SVM (RBF, C = 1000, γ = 0.01)	0.70	0.70	0.71	0.71
VGGFace-A	0.77	0.70	0.73	0.72
VGGFace-B	0.53	0.49	0.51	0.49
VGGFace-C	**0.84**	**0.76**	**0.79**	**0.79**

4.1 Experiments with ML-Face and ML-Body

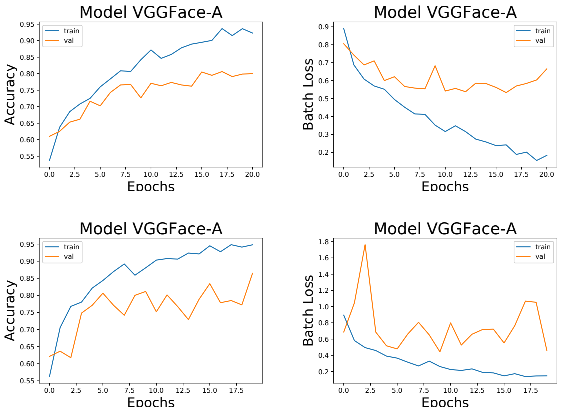
We trained these models with features extracted from their respective faces and bodies of Mary and Gabriel. For each of these models, we did a grid search for finding the best parameter for tuning the performance of the models. Table 2 and Table 3 show the results of these methods for face and body features, respectively. We can see that their performance is not consistent between both the features. In general, the performance is better with face features in comparison to body features. Furthermore, we see that for face features as well as for body features, SVMs show the highest accuracies. The drop in performance of these methods, when face features are compared to body features, indicate that they are unable to process data with higher visual complexity (or contextual content).

4.2 Experiments with VGGFace-* Models

All of these models are pretrained VGGFace [36] models (pretrained ResNet50 on the VGGFace dataset), hence we use a low learning rate of $1e-4$ with batch size of 32 for fine-tuning. Training is done on a randomly generated split of training and validation set [5] of images from our dataset of Mary and Gabriel.

VGGFace-A:. Table 3 shows that VGGFace-A outperforms all the traditional ML methods for all the metrics, when trained on body images. It gives comparable performance in comparison to SVM with RBF kernels. Figure 5 (row 1) shows the accuracy and loss plots during the training of the model. We can see that the training accuracy increases consistently while the validation accuracy stagnates at some point. The model training is stopped when the validation loss stops to improve. We use a tolerance of 0.05 for validation loss. If the validation loss is lower than the tolerance value for at least 10 epochs, the model training is halted. We observed empirically that this avoids overfitting the model. The confusion matrix for the model is shown in Figure 6 (a), where we can see that the model incorrectly predicts 56/200 test samples for both the classes giving a good quantitative performance.

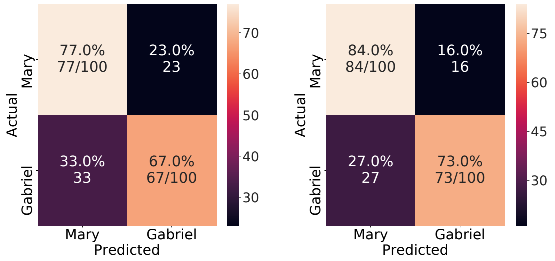
VGGFace-B:. This model is a special case, where we fine-tune the VGGFace (ResNet50) directly on the styled dataset. The styled dataset is a combination of images of people (males and females, thereby providing the context for training a two-class classification model) and the styles of the annunciation scenes transferred to those images. Table 3 shows the performance metrics of this model



(a) Accuracy Curves

(b) Loss Curves

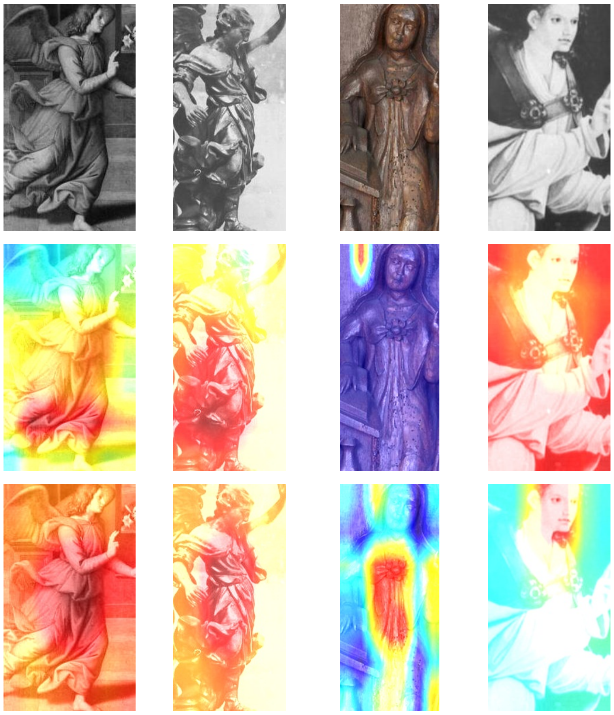
Figure 5: Training and Validation loss and accuracy curves for models VGGFace-A and VGGFace-C.



(a) VGGFace-A

(b) VGGFace-C

Figure 6: Confusion Matrices on test set of images of Mary and Gabriel.



(a) (b) (c) (d)

Figure 7: *Columns (a, b)* are positively predicted samples of Gabriel from the test set. *Columns (c, d)* are positively predicted samples of Mary from the test set. *Row 1* shows the original images, *Row 2* shows the CAMs generated for model VGGFace-A, whereas *Row 3* shows the CAMs generated for model VGGFace-C.

on the test set of images of Mary and Gabriel, which is inferior to that of VGGFace-A. This makes sense since the model has not seen the actual images of Mary and Gabriel, only their styles. When tested on the test set of styled images, the performance metrics are: *Pr: 0.73, Re: 0.82, F1: 0.77, Acc: 0.76*, which are even better than using VGGFace-A (Table 3 (row 5)). These metrics suggest that the model has adapted quite well on the styled data due to their larger database size.

VGGFace-C:. Motivated by the performance of the previous models, we take *VGGFace-B* as our base model and fine-tune it on the Mary and Gabriel dataset. We take the standard parameters and train it until the validation loss saturates. Figure 5 (row 2) shows the accuracy and loss progression over the training epochs for the model. Similar to VGGFace-A, we stop the training when the validation loss does not improve. Here as well we consider the tolerance value for validation loss to be 0.05, so if the validation loss is within this tolerance value for at least 10 epochs, then the training is stopped to avoid overfitting. Table 3 shows that this model outperforms all other models for all the metrics. Figure 6 (b), shows the confusion matrix for this model, we see that the model misclassified 43/200 samples in contrast to 56/200 samples in case of VGGFace-A. We can conclude that the network is able to make use of the domain knowledge acquired through the training on the styled images.

4.3 CAM Analysis on CNN models

Visualization of the deep networks has become one of the most important and interesting aspects of deep learning, mainly to understand the internal working of the networks [49]. Deep CNN networks are powerful enough to learn the localization of objects in higher layers [52]. *Class Activation Maps* (CAMs) [52] are an interesting way to visualize the image regions where the deep network is focusing while making the predictions about it's class.

Figure 7 shows the CAMs for positively predicted test samples of Mary and Gabriel for VGGFace-A (row 2) and VGGFace-C (row 3) models. We can see how the network tries to localize at the body (including the dress), wings and the neighboring information while predicting for Gabriel and Mary. The results from the CAMs strengthen our argument as to why VGGFace-C is a better model as compared to VGGFace-A. For Figure 7(a, b), VGGFace-A uses features located at the lower part of the clothing, whereas VGGFace-C perceives the whole body. This also shows that the use of the whole body for recognizing character gives more context for the model to learn about

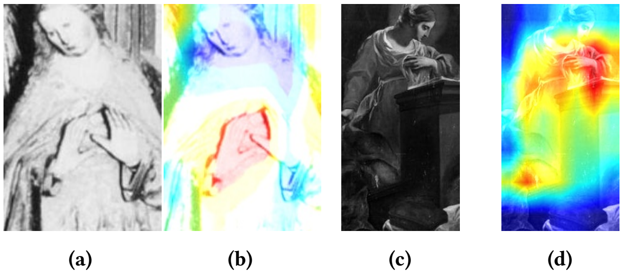


Figure 8: (a) and (c) are the original images of Mary from the test set and *(b), (d)* are their corresponding CAMs.

the character. In Figure 7(c) VGGFace-C again looks at the body part of Mary for making the prediction, whereas for Figure 7(d) it focuses on the facial area, as opposed to VGGFace-A which looks at other regions of the images except the face. Interesting to note here that both the models are making the correct predictions by looking at different regions.

A very peculiar information that the network tries to capture is observed in Figure 8, where it has correctly classified the class as Mary, but in this case, the focus of the network is on the hand positions of Mary. This is a very useful semantic information that the network is trying to capture when it is provided domain related knowledge.

5 DISCUSSIONS AND CONCLUSION

We demonstrated that the traditional ML techniques are insufficient to learn complex and more contextual features present in the bodies of the characters. Specifically for images from artworks, it is important to visualize more context (body as opposed to only face) to allow for a better analysis of the characters. We showed that styles from one domain can be transferred to another, and this information from styled images is beneficial for improving the performance of the deep CNN models. By looking at the CAMs, we were able to see how these networks capture semantic information present in the annunciation scenes to make the predictions.

Recognizing characters in art history is a complex problem given the diversity of ways in which the protagonists can be described. We demonstrated that deep CNN models are able to learn the required domain knowledge for recognizing character using style-transferred images. This technique allows that available datasets can be style-transferred and then used to fine-tune models before training on the actual datasets, thereby reducing excessive reliance on larger datasets in art history. In the end, it is important to note that using the whole body annotations of Mary and Gabriel, the models are able to perform better since they are able to capture more contextual and semantic information.

ACKNOWLEDGMENTS

This work was funded by Emerging Fields Project *ICONOGRAPHICS* within the Framework of the FAU[2] Emerging Fields Initiative. This work has also been partly supported by the Cross-border Cooperation Program Czech Republic – Free State of Bavaria ETS Objective 2014-2020 (project no. 211). We would also like to thank NVIDIA for their generous hardware donations.

REFERENCES

[1] August 20, 2019. The Annunciation. https://www.slam.org/collection/objects/7082/

[2] August 20, 2019. Verkündigung Mariae zwischen den hll. Sebastian und Lucia. https://www.sammlung.pinakothek.de/de/artist/mariotto-albertinelli/verkuendigung-mariae-zwischen-den-hll-sebastian-und-lucia

[3] Shumeet Baluja and Henry A Rowley. 2007. Boosting sex identification performance. *International Journal of computer vision* 71, 1 (2007), 111–119.

[4] Peter Bell, Joseph Schlecht, and Björn Ommer. 2013. Nonverbal communication in medieval illustrations revisited by computer vision and art history. *Visual Resources* 29, 1-2 (2013), 26–37.

[5] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. 2011. Unsupervised metric learning for face identification in TV video. In *2011 International Conference on Computer Vision*. IEEE, 1559–1566.

[7] Elliot Crowley and Andrew Zisserman. 2014. The State of the Art: Object Retrieval in Paintings using Discriminative Regions.. In *BMVC*.

[8] Elliot J Crowley, Omkar M Parkhi, and Andrew Zisserman. 2015. Face Painting: querying art with photos.. In *BMVC*. 65–1.

[9] Elliot J Crowley and Andrew Zisserman. 2014. In search of art. In *European Conference on Computer Vision*. Springer, 54–70.

[10] Leonardo da Vinci and Frank Zöllner. 2003. *Leonardo da Vinci: 1452-1519: sämtliche Gemälde und Zeichnungen*. Taschen.

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

[12] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. *arXiv preprint arXiv:1904.10699* (2019).

[13] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179.

[14] Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin, and Chinchilla Doudou. 2014. Learning deep face representation. *arXiv preprint arXiv:1403.2802* (2014).

[15] L. Fei-Fei and P. Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. 524–531 vol. 2. https://doi.org/10.1109/CVPR.2005.16

[16] Noa Garcia and George Vogiatzis. 2018. How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[17] Vittoria Garibaldi. 2004. *Perugino*. Vol. 197. Giunti Editore.

[18] Beatrice A Golomb, David T Lawrence, and Terrence J Sejnowski. 1990. Sexnet: A neural network identifies sex from human faces.. In *NIPS*, Vol. 1. 2.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Oskar Holl. 1968. Engel. *Lexikon der christlichen Ikonographie* 1 (1968), 626–642.

[21] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[22] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[23] Leonardo Impett and Franco Moretti. 2017. *Totentanz. Operationalizing Aby Warburg's Pathosformeln.* Technical Report. Stanford Literary Lab.

[24] Theodor Klauser. 1962. Engel X (in der Kunst). *Reallexikon für Antike und Christentum* 5 (1962), 258–322.

[25] Heinrich Krauss. 2001. *Kleines Lexikon der Engel: von Ariel bis Zebaoth.*

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[27] Lehre. [n. d.]. The prometheus Image Archive: High-quality images from the fields of arts, culture and history. https://www.prometheus-bildarchiv.de/

[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8759–8768.

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[31] David G Lowe et al. 1999. Object recognition from local scale-invariant features.. In *iccv*, Vol. 99. 1150–1157.

[32] Paolo de Matteis. 1712. The Annunciation.

[33] Johann Michel. 1962. Engel VI (Gabriel). *Reallexikon für Antike und Christentum* 5 (1962), 239–243.

[34] Baback Moghaddam and Ming-Hsuan Yang. 2002. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (2002), 707–711.

[35] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition.. In *bmvc*, Vol. 1. 6.

[37] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 121–135.

[38] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv* (2018).

[39] Alfons Rosenberg. 1967. *Engel und Dämonen, Gestaltwandel eines Urbildes*.

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[41] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)* (July 2016).

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (2014), 1573–1585.

[43] Josef Sivic, Mark Everingham, and Andrew Zisserman. 2005. Person spotting: video shot retrieval for face sets. In *International conference on image and video retrieval*. Springer, 226–236.

[44] Josef Sivic, Mark Everingham, and Andrew Zisserman. 2009. "Who are you?"-Learning person specific classifiers from video. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1145–1152.

[45] Gjorgji Strezoski and Marcel Worring. 2017. OmniArt: Multi-task Deep Learning for Artistic Data Analysis. (2017). arXiv:1708.00684 http://arxiv.org/abs/1708.00684

[46] Georges Tavard. 1982. Engel V. Kirchengeschichtlich. *Theologische Realenzyklopädie* 9 (1982), 599–609.

[47] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. 2019. Feature Transfer Learning for Face Recognition With Under-Represented Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Nadir Yurtoğlu. 2018. http://www.historystudies.net/dergi//birinci-dunya-savasinda-bir-asayis-sorunu-sebinkarahisar-ermeni-isyani20181092a4a8f.pdf. *History Studies International Journal of History* 10, 7 (2018), 241–264. https://doi.org/10.9737/hist.2018.658

[49] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[50] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. 2019. Unequal-Training for Deep Face Recognition With Long-Tailed Noisy Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856* (2014).

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127, 3 (2019), 302–321.