UNIVERSITY OF TARTU

Institute of Computer Science

Computer Science Curriculum

Henri Harri Laiho

Recognition as Navigation in Energy-Based Models

Bachelor's Thesis (9 ECTS)

Supervisors:

Raul Vicente Zafra, PhD

Jaan Aru, PhD

Tarun Khajuria, MSc

Tartu 2021

Recognition as Navigation in Energy-Based Models

Abstract:

Human vision has an exceptional ability to recognize complex signals from limited and ambiguous observations, which is believed to comprise lower-level processes generating possible explanations for the observations, and higher-level systems selecting the most plausible ones of them. There is a lack of comparable mechanisms in modern artificial intelligence visual recognition solutions that would enable an improved generalization and robustness. This thesis proposes and studies a novel brain-inspired algorithm for face recognition which tackles the problem from a new angle – recognition can be solved as a navigation problem in a space of latent representations. Further, we show that the steps of this navigation correspond to sensible images that the model "imagines" during the process of navigation, comparable to a human imagining possible explanations to the observations which he/she is trying to recognize as an object or a person. In addition to this, we present that with some parameter tuning the algorithm can improve the separability of correct and incorrect navigation trajectories – like the explanations proposed by lower-level processes in the brain – as Fisher's discriminant ratio by up to 0.14 which, according to our guess, corresponds to an increase in accuracy between 5-15%.

Keywords:

face recognition, navigation, energy-based models, latent representation, vision

CERCS: P176, Artificial intelligence

Tuvastamine kui tee leidmine energiapõhiste mudelite abil

Lühikokkuvõte:

Inimeste nägemisvõime suudab tuvastada keerulisi mustreid vikerkesta kaudu saabuvatest mitmestest vaatlustest. Usutakse, et meie nägemine koosneb madalatasemelistest protsessidest, mis pakuvad välja seletusi vaatlustele, ning kõrgematasemelistest süsteemidest, mis valivad neist kõige usutavama seletuse. Võrreldavaid üldistamisvõimet ja töökindlust parandavaid mehhanisme pole kasutusel selle töö kaasaegsetes pildituvastuse tehisintellekti lahendustes. See bakalaureusetöö uurib ning pakub välja uudse ajust inspireeritud näotuvastusalgoritmi, mis läheneb ülesandele uue nurga alt: tuvastamist saab vaadelda kui teeleidmisülesannet varjatud esituste vektorruumis. Me näitame lisaks, et tee leidmisel läbitud sammud vastavad mõistlikele kujutistele, mida meie kasutatud mudel teatud mõttes "kujutab ette" teeleidmisprotsessi käigus, ning mis on võrreldavad inimese kujutluspiltidega, mida aju välja pakub vikerkestale saabuva pildi seletuseks. Peale selle me esitame tulemused, mis näitavad, et mõningase algoritmi parameetrite tuunimise järel suudab algoritm suurendada õigete ja valede leitud teede – ajus alateadvuse väljapakutavate seletuste – eristatavust Fisheri diskriminandi suhte kujul kuni 0.14 võrra, mis meie arvates võib vastata mudeli täpsuse kasvule 5-15%.

Võtmesõnad:

näotuvastus, navigatsioon, energiapõhised mudelid, varjatud esitus, nägemine

CERCS: P176, Tehisintellekt

Table of Contents

1. Introduction 5

2. Background 7

  2.1. Problem of Recognition 7

  2.2. Recognition in AI 7

  2.3. Recognition in the Brain 8

3. Methodology 9

  3.1. Setup of the Experiment 9

    Dataset 9

    Task – Recognition of Identity 9

  3.2. Models 11

    Energy-Based Models 11

    Origin 11

    Architecture 11

  3.3. Method 13

  3.4. Evaluation 15

4. Results 17

  4.1. Visual Analysis 17

    Energy Landscape and Placement of Identities 17

    Embedded Space Walks 18

    Parameter Tuning 20

    Visual Approximations of the Embedded Space Walk Steps 22

  4.2. Objective Assessment on Method Performance 23

5. Discussion 26

6. Conclusion 27

7. References 28

Appendix 30

  I. License 30

  II. Visual Approximations of Walks 31

  III. Visual Approximations of Walks 32

  IV. Visual Approximations of Walks 33

  V. Parameters for the Best Results 34

1. Introduction

Machine learning models need to achieve an exceptional ability to generalize in order to be reliably applied on complex tasks like autonomous driving, strategic decision-making and eventually even software development. In many modern AI models there are the issues of models being deceivable by adversarial samples and limited generalization to cope with biases in out-of-distribution data [1], [2]. Yet, humans can perform most of such tasks without these problems, therefore a solution to the generalization problem must have been figured out by the brain. This fact makes it beneficial to study the brain and develop AI algorithms that are drawn from its workings.

In this thesis we explore a new idea for completing a complex task in a way that is plausibly similar to how the brain solves the task. By taking inspiration from the brain, there is a preconception that this method would be robust and generalizing. Another ambition is to get insight into how the brain may perform recognition and other similar tasks, to which this method can be conveyed.

We propose a new approach to the task of recognition like face recognition in smart device screen locks or labelling of faces in photographs. This specific task was selected for its complexity, availability of data and pretrained models, and the interest in how the brain solves the task. This method is potentially also expandable to a larger variety of problems from traffic sign detection to plant species recognition. The aim of this thesis is to implement, evaluate and analyse our proposed method for face recognition that solves the problem as a problem of navigation.

There is indication that while performing certain recognition tasks, the brain tries to compare the subject of recognition to categories it is familiar with [3]. We believe this comparison is implementable as navigation in a space of imaginations, where every form of reachability indicates its respective type of similarity between the comparables. The idea we explored translates the problem of recognition to navigation in a hidden vector space of an image processing neural network model. In contrast to one of the best-performing models to date – FaceNet by Google – our method utilizes navigation in a hidden space instead of just comparing the hidden space representations by their initial position before the navigation step [4]. We test the method on the dataset of CelebA[1], and the pretrained models from [5]. The use of our method slightly improved the clustering of identities in the hidden space, however insignificant in the context of modern methods.

This work consists of three main chapters – background, methodology and results. In the background we explain the task of face recognition, analyse how this is being solved by current AI models and establish that these models do not yet achieve human-level generalization and therefore require further research and development. The chapter of methodology describes in high detail how the aim of the thesis was achieved by first

[1] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

reformalizing the face recognition task to suit the navigation approach, secondly, outlining the structure of the image processing model, thirdly, formally specifying the method itself, and finally providing an overview of the techniques and metrics used to measure and evaluate the method. The chapter of results presents thorough visual analysis of the method and summarizes it with concrete metrics on the performance of the method.

2. Background

Even though there has been a lot of research on the problem of recognition, it is still not known, how recognition in the brain works. In this section we discuss some related work on the topic, first by interpreting the problem, then mapping some present AI solutions. Finally we hypothesize how the problem is solved by the brain with the support of related work and in the end converge on the idea that we explore in this thesis.

2.1. Problem of Recognition

The task of recognition appears to be trivial as, for example, humans are able to recognize objects in a new scene within seconds. However some complications arise when implementing recognition in AI and even more when trying to understand how it is done in the brain. Generally, recognition can be seen as a classification task, where raw inputs received from the world are discriminated into a discrete class that the inputs represent. [3] The inputs from the world can be distorted or severely compressed. In visual recognition the input is a 2D projection on the retinae of a 3D object that is being recognized. This projection introduces a large amount of ambiguity to the input, which in the brain is disambiguated by higher-level cognitive processes via aggregating the input with other signals from the world [6].

2.2. Recognition in AI

In face recognition and species identification models a solution to the problem has been to execute feature extraction on the input, shrinking the dimensionality of the data, and thereafter to apply a conventional classifier to determine the class, or recognize the object [4], [7].

Current state-of-the-art human face recognition AI models have achieved accuracies over 95%, in some cases even over 99% [4], [8]. This good performance makes it seem that the problem has been practically solved. The FaceNet model is shown to be adversarially robust and encodes images into a 128-dimensional vector representation where the euclidean distance between two vectors is proportional to the similarity of the inputs [4]. This indicates that AI models have achieved near human-level performance in face recognition on these benchmarks and the way how these models work with embeddings and feature extraction appears also similar to how the brain performs the task. It however, does not imply that these models reproduce the recognition process in the brain. Additionally, it is known that some conventional convolutional neural network models tend to predict too much based on the texture rather than shape of the objects – an issue that does not reproduce in human vision [9]. Therefore there is motivation to explore other ideas to solve the task in more abstract ways.

2.3. Recognition in the Brain

Humans associate objects with hidden compressed representations – subjective meanings, concepts or notions [10, p. 135]. It is plausible that in the process of recognition the brain does feature extraction on input data to infer an abstract concept, similar to AI recognition models like in [4]. After this step it does not seem plausible that the brain calculates euclidean distance of two concepts or directly performs kNN classifications like in FaceNet. One option is that some navigation in the concept space, driven by the activity of the recurrent neuron populations, is done.

There is a lot of recent literature relating navigation to planning and other thinking tasks, which hints that navigation has an important role in cognitive processes [11], [12]. However there appears not to be any material on solving recognition in this way, which is why this thesis will be useful. One example that demonstrates the usefulness of the concept of navigation in recognition comes from the classic task of "mental rotation". When deciding whether two 3D shapes displayed under a different angle have the same shape, the brain solves the task with linear complexity with respect to the angle of rotation between the two shapes. A logical explanation to this is that the brain mentally rotates the shape to decrease the angle between the shapes. [13, pp. 35–36] This observation suggests that some form of navigation can be used to solve a recognition task. However the navigation is strictly limited to shape-preserving transformations like rotation, which brings a need for, preferably automatically, determining such allowed transformations that can be applied during the navigation process. In the latent space of generative adversarial networks it is possible to learn transformations that correspond to any specified transformation in the output space [14]. This method would enable one to determine these transformations in a hidden space, but the allowed transformations would still have to be manually defined.

Another approach would be constraining the navigation via the use of energy landscapes of energy-based models (EBMs), like Hopfield neural networks, so that movement is allowed only in directions that decrease the energy. Hopfield neural networks were inspired by the brain and the method of how neurons store data as memories [15]. Therefore, it makes sense to implement navigation between memories as movement between states of an EBM. It has been shown that EBMs have great out-of-distribution generalization and can be used for data generation as well as classification with robustness toward adversarial samples [16]. Like Hopfield neural networks, these EBMs can be used to query data stored in the model by providing partial input and using an energy-minimizing update rule to converge to a stable state [15]. It is not known if or how the energy landscape helps to find allowed identity-preserving transformations, so another objective of this thesis is to explore and hopefully answer this question.

3. Methodology

Since our goal was to design a brain-inspired algorithm for the task of recognition, we decided to narrow the task to human face recognition, which is a well studied field with enough data and specific pretrained models available. After setting up the experiment with data and a formal task to solve, the pretrained models were rewired to enable navigation in an extracted feature space or an embedded space. As the last step, the navigation algorithm was designed and evaluated.

3.1. Setup of the Experiment

In order to know how to design and, in the end, evaluate an AI solution one needs a dataset and a task to solve. This section describes the dataset and task for which we implemented the method.

Dataset

The dataset used in this experiment is called CelebFaces Attributes Dataset (CelebA)[2]. It contains 202599 face images of 10177 celebrities. The images have been annotated with 40 binary attributes. The dataset provides a wide range of diversity in terms of background, color, facial expressions, age, face angle [17]. The images were originally in the resolution of 218 x 178 pixels, but were then rescaled into the resolution of 128 x 128 pixels for compatibility with the pretrained model.

The fact that the dataset has been annotated with identity information and contains about 20 samples per identity on average, makes it suitable for the task of identity recognition, or more specifically, for evaluation of the novel approach to the problem that was studied in this work.

Task – Recognition of Identity

We aim to develop a brain-inspired solution to the following task: for given images X_A and X_B, predict if the objects on images X_A and X_B share a common property or not. Depending on the underlying data, the property may be defined in multiple ways. For example in case of photos of animals or plants, the property of interest might be the species of the being. In case of photos of a single species, like the dataset CelebA, which contains only images of humans, a property worth analysing would be the identity of the person. By defining the task in this way, as binary classification, opposed to just predicting the property on a single image which would be multiclass classification, we make the solvable in a way that can handle novel out-of-distribution values of the property, for example if a new species is discovered or new identities are added to the set of known identities. Otherwise, unknown classes would be predicted as the class "other", but this way the new classes can be added by memorizing at

[2] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

least one sample of the class as X_B; and then one can query if a sample belongs to the class by providing X_A as a parameter. Another reason for this formulation comes from that it is plausibly more similar to how the brain solves the task of recognition. Due to the specifics of the method, it was not possible to formulate the task as kNN classification like in [4].

In the scope of this thesis, the discussed method was studied on a dataset of human portraits (CelebA) and the property of interest was identity of the person. In this case, the task is almost identical to the passport task which is solved at state border checkpoints when comparing the image on the passport to the document holder's face. More formally, the task to be solved was "for given portraits of people X_A and X_B, predict if the identity of the person in X_A equals to the identity of the person in X_B". Identity of the person is invariant over varying, for example, hair color, glasses, accessories, age, facial expressions, hair shape, and more, but (usually) variant over varying shape of the face, sex, race, position of mouth, eyes and nose, and more. Figure 1 shows 4 examples of the given image pairs (X_A, X_B) and their respective expected labels (Y).

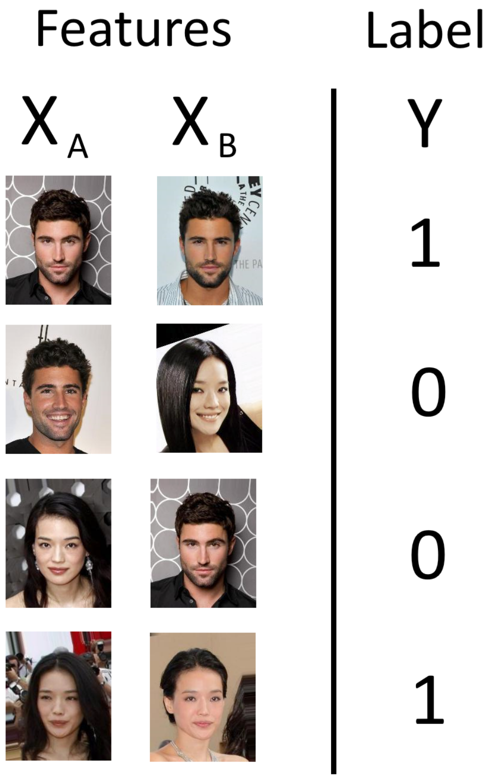

Figure 1. 4 example samples of data used in the experiment. There are images of one female and one male. The label is 1 if and only if both images depict the same person, 0 otherwise.

The original CelebA dataset contains single images, but the task requires pairs. In order to suit the task, a data loader was defined in the pipeline which yields pairs of CelebA images with respective annotations and labels. By taking random pairs of images from the dataset, the frequency of label 0 would be much greater compared to the frequency of label 1. To mitigate this issue and balance the dataset, the data loader was modified to provide pairs of images of common identity with probability 0.5.

3.2. Models

As proposed by [6], the process of vision in the brain consists of a generative and a discriminative system. Therefore we propose solving the task with the use of energy-based models (EBM), that appear to hold both of these properties, as discussed in the following section. Additionally, in this section we describe the modifications to the models needed to implement the method that we study in this thesis.

Energy-Based Models

EBMs map every configuration of inputs to a scalar value, which intuitively represents the compatibility between the inputs, or simplified in the context of this thesis – a value proportional to the probability of finding the input configuration in the dataset [18].

It has been shown that EBMs can be trained and used to generate images. This is done by initializing the input with random value and then learning the input with gradient descent, using the energy as a loss term. Such EBMs can be trained using a technique called contrastive divergence. While applying contrastive divergence, the energy landscape is formed by raising the energy at input configurations of generated fake images and lowering energy at input configurations of real dataset images. [16]

In this application, the EBMs are similar to the discriminators in generative adversarial networks with the difference that they do not not solve a classification task, but a regression task. This allows the use of the model output directly as a loss term for gradient descent to optimize the model input as to minimize (or maximize) the model output [16]. This in turn enables the use of EBMs as generative models in addition to the discriminative process of energy prediction.

Origin

The method that was studied requires an EBM which has been trained on similar data. Due to time limitations, pretrained models, which set constraints on the structure of the model, were used. The models had been designed and trained by Du et al., 2021, and included conditional EBMs for male, old, smiling and wavy-haired celebrity images [5]. As an effect of this thesis, the trained models were also made publicly available[3].

Architecture

The models consist of 3 parallel convolutional networks, each with different amounts of downsampling in the input. With the aim to reduce the dimensionality of the embedded

[3] https://github.com/yilundu/improved_contrastive_divergence

space, the following variants of the embedded space were experimented with: 4x downsampled submodel split before the final dense energy map layer (a), the 4x downsampled 2 residual blocks before the energy map layer (b), and all levels of downsampling split directly before the energy map layer (c). See figure 3 for a visualization.

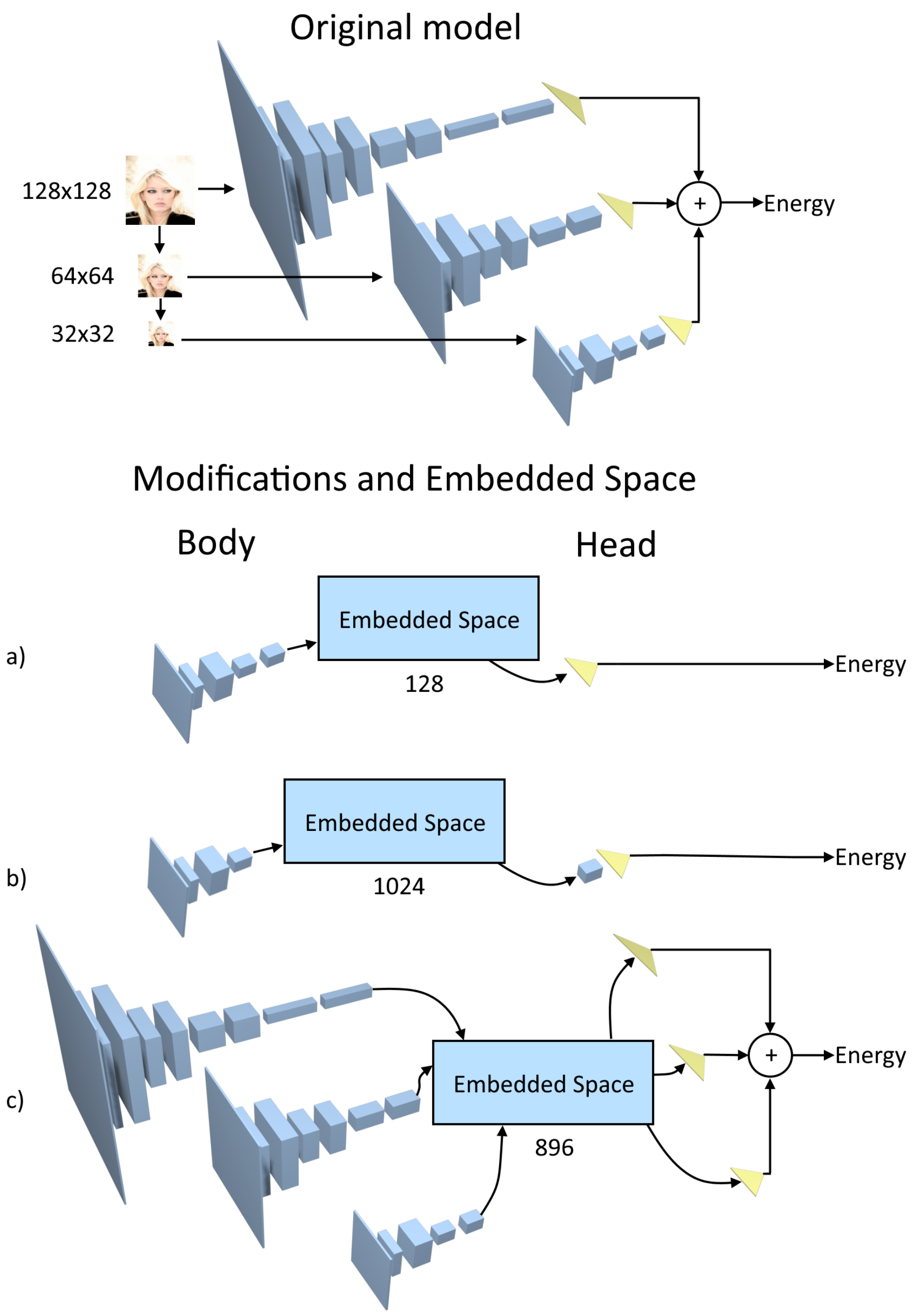

Figure 3. Structure of the models. Each model has 3 submodels, the input of 2 of the submodels is scaled down respectively to the resolutions 64x64 and 32x32 pixels. The output of the submodels is a scalar per every input image. The final energy is calculated as a sum of the outputs of the submodels. 3 variations of an embedded space with dimensions 128 (a), 1024 (b) and 896 (c) were defined as shown. We call the part of the model that transforms images to embedded space vectors the *Body* and

the part that transforms embedded space vectors to energy, the *Head*. In case of variation c, both *Body* and *Head* contain 3 submodels and the embedded vectors are obtained by concatenating the inputs. In the *Head* part, the vector is split back into 3 vectors to reverse the concatenation.

Vectors in the embedded space variations (ESV) a, b and c will have accordingly the sizes of 128, 1024 and 896 (see Figure 3). The body of the model, *Body(x)*, converts the image *x* into an embedded space representation which is a vector of a size depending on how the embedded space is defined. The head of the model, *Head(z)*, converts the embedded space vector *z* into a scalar energy. The sub-models Body and Head were defined so that the following equation holds: *Model(x) = Head(Body(x))*, where *Model(x)* is the original EBM.

3.3. Method

We make the assumption that the embedded space representations of pairs with common identity are more easily reachable from one another in the context that moving uphill on the energy landscape is difficult. The correctness of this assumption is studied in this thesis.

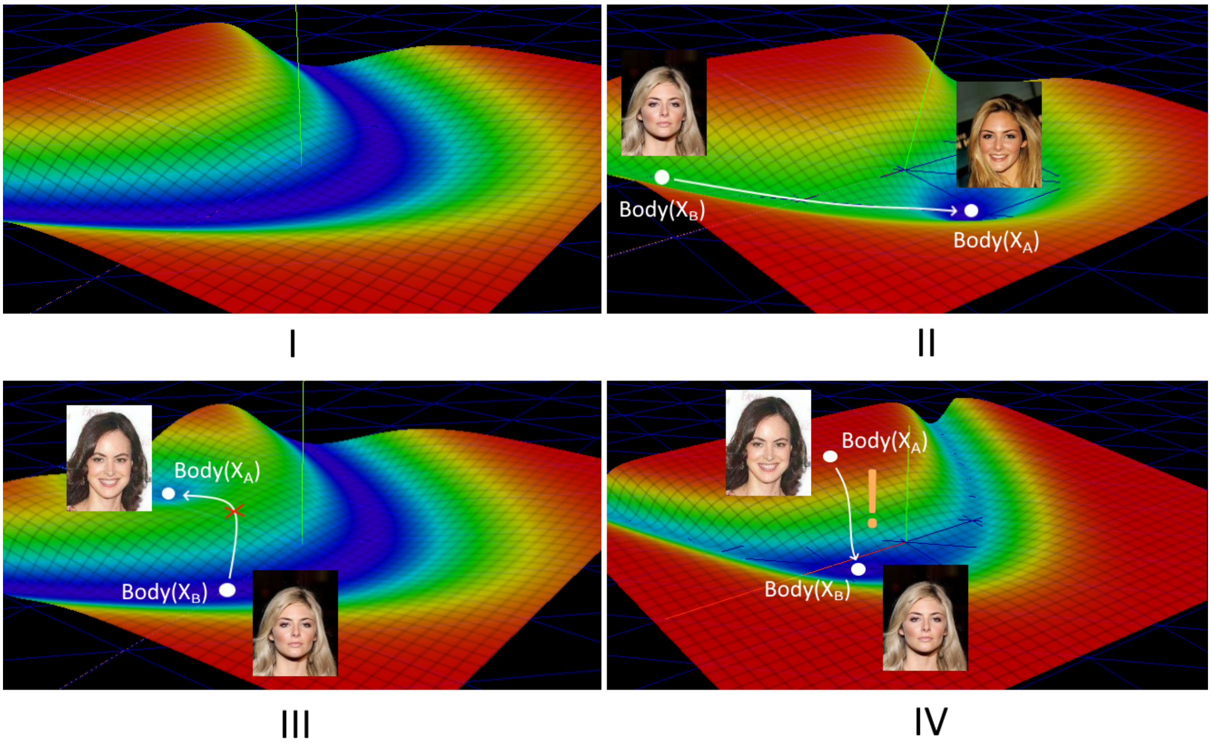

Figure 4. (I) Hypothetical energy landscape of an EBM, where the horizontal axis are inputs and the vertical axis represents the output, the energy. (II) Embedded representations of photos X_A and X_B of a single person. An energy valley is generated around one of the points so that weak obstacles can be overcome with a gradient descent walk. (III) Photos of different people are less reachable according to our assumption, even if the energy valley is present. Here *Body(X_A)* is unreachable from *Body(X_B)* due to the peak between. (IV) If we would query reachability in the other direction we would get the opposite result, therefore the prediction, needs to be the conjunction of reachability from *Body(X_A)* to *Body(X_B)* and reachability from *Body(X_B)* to *Body(X_A)*.

We propose the following solution to the task. First use the bodies of the models to obtain embedded space representations of the images, $Z_A = Body(X_A)$ and $Z_B = Body(X_B)$. Then add a gaussian energy valley around Z_B: $E_{\beta,\sigma}(z) = Head(z) - \beta e^{\frac{-\Sigma[(z-Z_B)^2]}{\sigma}}$

Use a descending gradient of Z_A w.r.t. $E_{\beta,\sigma}(Z_A)$ to update Z_A for n steps. Calculate a distance measure between the latest z and Z_B. Repeat the same for Z_B. Output prediction based on the distance measures between the latest Z_A and initial Z_B, and between latest Z_B and initial Z_A. We propose taking the maximum of the 2 distances since according to our assumption, different identity should be predicted in case there is no reachability in one direction. See Algorithm 1 for details.

Algorithm 1. Embedded Space Walk Algorithm

```
Input: $X_A$, $X_B$; Parameters: number of steps $n$, energy valley width $\sigma$, energy valley
       depth $\beta$, walking step size $\alpha$, distance measure $D(z_1, z_2)$

$Z_{A,0} \leftarrow Body(X_A)$; $Z_{B,0} \leftarrow Body(X_B)$
for $i = 1, 2, ..., n$ do:
    $Z_{A,i} \leftarrow Z_{A,i-1} - \alpha\nabla_{Z_{A,i-1}} E_{\beta,\sigma}(Z_{A,i-1})$,
        $where\ E_{\beta,\sigma}(z) = Head(z) - \beta e^{\frac{-\Sigma[(z-Z_{B,0})^2]}{\sigma}}$
end for
for $i = 1, 2, ..., n$ do:
    $Z_{B,i} \leftarrow Z_{B,i-1} - \alpha\nabla_{Z_{B,i-1}} E_{\beta,\sigma}(Z_{B,i-1})$,
        $where\ E_{\beta,\sigma}(z) = Head(z) - \beta e^{\frac{-\Sigma[(z-Z_{A,0})^2]}{\sigma}}$
end for
return $max(D(Z_{A,0}, Z_{B,n}), D(Z_{B,0}, Z_{A,n}))$
```

Additionally, gaussian noise was added at every step of the walk in the update rules of $Z_{A,i}$ and $Z_{B,i}$, a parameter was added to control the strength of the noise. There are 4 models that were used in various combinations to implement this method. In case of using multiple models simultaneously, the mean of the maxima of the final distances was calculated. By combining multiple models in this way, the solution is expected to generalize better than when using a single model. The final predictions were obtained by predicting the label which, in the training set, had a mean distance nearest to the sample's distance (see decision boundary on Figure 5).

On the assumption that images of people with the same identity are placed in common energy valleys and are easily reachable from one another when some attraction is applied, Algorithm 1 will return a low value for pairs with common identity and a high value for pairs with different identity, given that the algorithm parameters α, β, σ, n and D are correct.

3.4. Evaluation

The embedded space energy landscape with respect to data placement was visualized on 2D plots, using PCA to find the most informative 2D view of the high-dimensional embedded space vectors. The same technique was used to visualize the embedded space walks, these plots were used to intelligently tune the parameters of the algorithm and visually evaluate the accuracy of the method on small test sets.

Additionally, to get deeper insight into if this method works, the walk steps on some model configurations were approximated back into the image space using a similar method to the Langevin sampling technique which is used for image generation on the same models as described in [5]. The difference here is that instead of energy, in this case the input image x was optimized w.r.t. the square error between *Body(x)* and z_{target}, where z_{target}, is the walk step being approximated. The optimization process was initialized with the image of the previous step.

The method was evaluated with accuracy and f1-score metrics and 2 custom metrics. To measure the metrics, first, pre-walk distances and post-walk distances (outputs of Algorithm 1) between pairs in the test set were calculated. Secondly, the means and standard deviations (STDEV) of these distances of positive and negative samples, before and after the walks, 4 distributions in total, were calculated (see Figure 5). Then for accuracy and f1-score a decision boundary was set between the mean of post-walk positives and the mean of post-walk negatives. The area under ROC could have been used as a metric, but since the labels were artificially balanced, for simplicity, labels were explicitly calculated for accuracy and f1-score.

One custom metric, called *rating*, describes how much did the walk algorithm improve the separability of the classes. It was calculated by subtracting the pre-walk Fisher's discriminant ratio (FDR) from post-walk FDR [19]. The other called, *FDR ratio*, describes how many times did the walk algorithm improve the class separability. If $\mu_{p,i}^2$, $\mu_{n,i}^2$, $\sigma_{p,i}^2$ and $\sigma_{n,i}^2$ are respectively positives mean, negatives mean, positives STDEV and negatives STDEV and if in case of *i=1* they are the post-walk statistics and in case of *i=2* pre-walk statistics, then the metrics would be calculated with the following formulas.

$$Rating = FDR_{post} - FDR_{pre} = \frac{(\mu_{p,1}-\mu_{n,1})^2}{\sigma_{p,1}^2+\sigma_{n,1}^2} - \frac{(\mu_{p,2}-\mu_{n,2})^2}{\sigma_{p,2}^2+\sigma_{n,2}^2}$$

$$FDR\,ratio = \frac{FDR_{post}}{FDR_{pre}} = \frac{(\mu_{p,1}-\mu_{n,1})^2}{\sigma_{p,1}^2+\sigma_{n,1}^2} \div \frac{(\mu_{p,2}-\mu_{n,2})^2}{\sigma_{p,2}^2+\sigma_{n,2}^2}$$

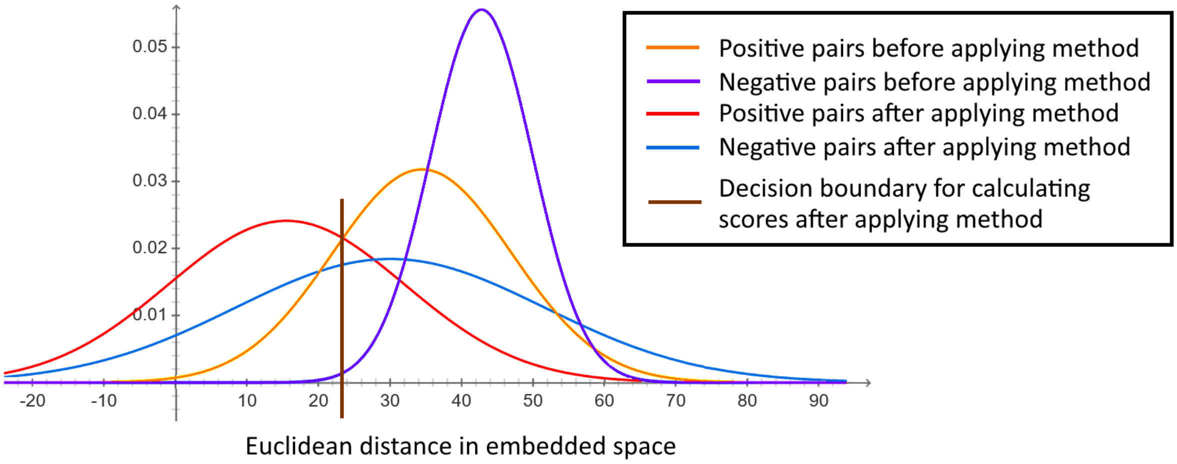

Figure 5. Method performance on a test set of 50 pairs. The bell curves show the distance between pairs of images in the embedded space. The yellow and purple distributions represent the data before running Algorithm 1. The red and blue distributions represent output of Algorithm 1. All distances in reality are nonnegative. The rating on this plot is -0.047, meaning that the FDR decreased by 0.047 due to running the algorithm.

4. Results

The code and links to the models that were used to run the experiments and produce the visualizations, are available on Github[4]. To convey the results, we first provide some visual insight into how the task is solved by the method. Then we present the model performance via the metrics described earlier.

4.1. Visual Analysis

As our goal was to determine if the energy function of the EBMs can be used to separate identities, the first idea was to visualize the energy function with respect to identities. Since this did not exhaustively confirm or disprove the effectiveness of the method, we present visualizations of the walk trajectories that show in which cases the algorithm works and in which cases it fails. We also demonstrate the effect of tuning the parameters of Algorithm 1 on the same plot type. Lastly, we show and analyse visual approximations of the walk steps to see the images that correspond to the embedded space vectors and better understand what is happening inside the models. All figures were made with the ESV *b* which was the most successful of the three, unless the ESV is specified near the figure.

Energy Landscape and Placement of Identities

A principal component analysis (PCA) was done on the embedded vectors of all images in the dataset of 3 randomly picked identities. Two principal components (PC) were used to plot the energy landscape and the embedded vectors (see Figure 6). The inverse transform of the PCA was used to render the energy landscape.

In Figure 6 we can see that the identities are relatively mixed in terms of placement on the two PCs. There also appear not to be any significant patterns of high energy separating the vague clusters of identities. Either too much information is lost when PCA transforms are done, or the assumption that identities can be separated by energy peaks is wrong.

[4] https://github.com/Henri-Laiho/ebm-recognition

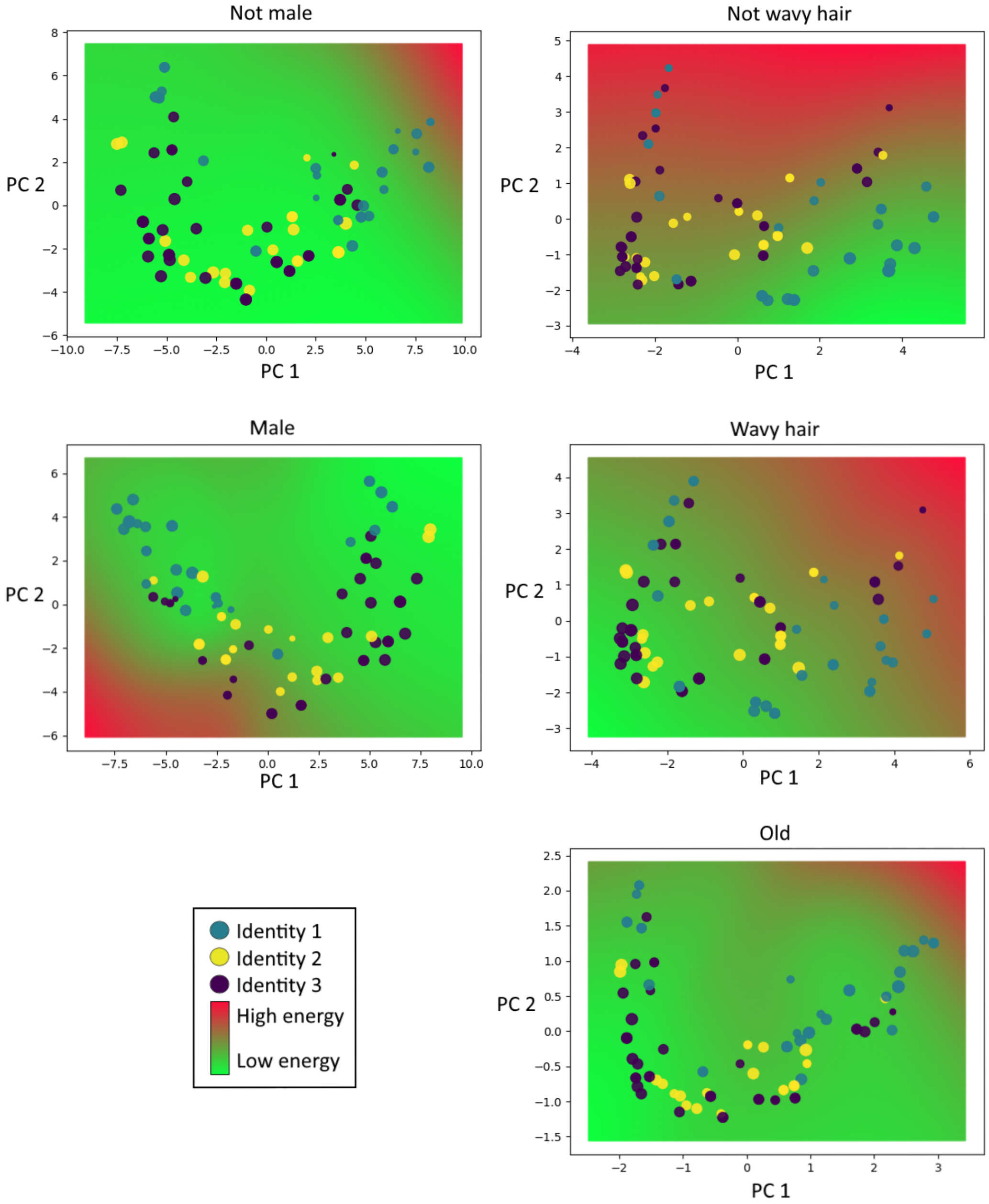

Figure 6. Placement of identities on the energy landscape of 5 model configurations. The axis of the plots correspond to the 2 principal components (PC) with the highest variance ratios. The background color represents the output of the model at the corresponding PC values. All samples in the dataset that have one of the three identities are shown.

Embedded Space Walks

In order to visually evaluate and see, in which direction should parameters be tuned, we plot walk trajectories – as in Figure 4 – but via the real steps $Z_{A,i}$, $Z_{B,i}$, $i \in \{1, 2, ... n\}$ that are calculated in Algorithm 1. Figure 7 provides an example of these plots with explanations. There are 3 positive pairs – the red, blue and cyan-green pairs – and 5 negative. The blue and red pairs are true positives, cyan-green is false negative, the pair of pink and yellow, and blue and yellow are false positives, the rest are true negatives, which means accuracy of 0.625.

Even though the distance between the starting-points of the red pair is high, the walk still connects them whereas the pair of pink and green, while at a similar distance from each other, does not become connected. The false positive pairs have the smallest distances between starting-points, which explains why the walk connects them – the attraction toward the target is greater near the target.

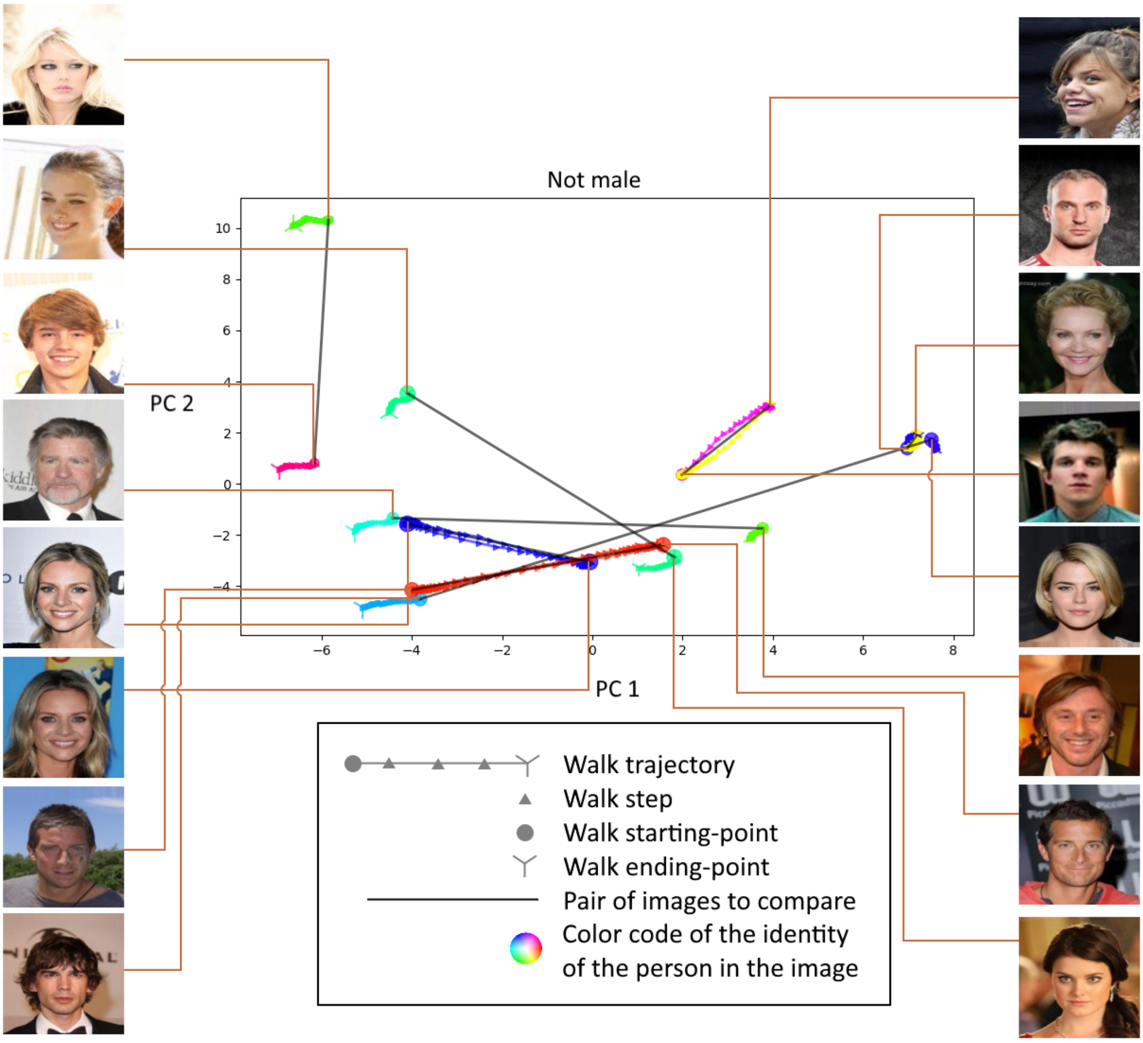

Figure 7. Explanation of the embedded space walk plot style. The plot shows a PCA view of the embedded space like the plots in Figure 6. Every walk starting-point represents one image in the embedded space as indicated by the orange lines. The main idea of the walks is to make pairs connected with black lines to reach each other via the walk trajectories, if they have the same identity color code and not connect if they have a different identity color code. The target point of every walk is the starting-point of its pair that is connected via the black line. A valley of attraction is added at the target point on the energy landscape that is being used to guide the walks.

The same test set of images in Figure 7 was used in Figure 8 as well as appendices II.-IV. As in the process of evaluation we use predictions made on the average distance given by an ensemble of models, we plot the walks on all models used in evaluation in Figure 8. In Figure 8 we can see that regardless of model, the walks behave very similarly. The model for males does not connect the red identity, which in fact is a male. The model for not wavy hair connects a negative pair of green and cyan identities. By taking the average over all 5 models, the mistakes of male and not wavy hair models will be corrected.

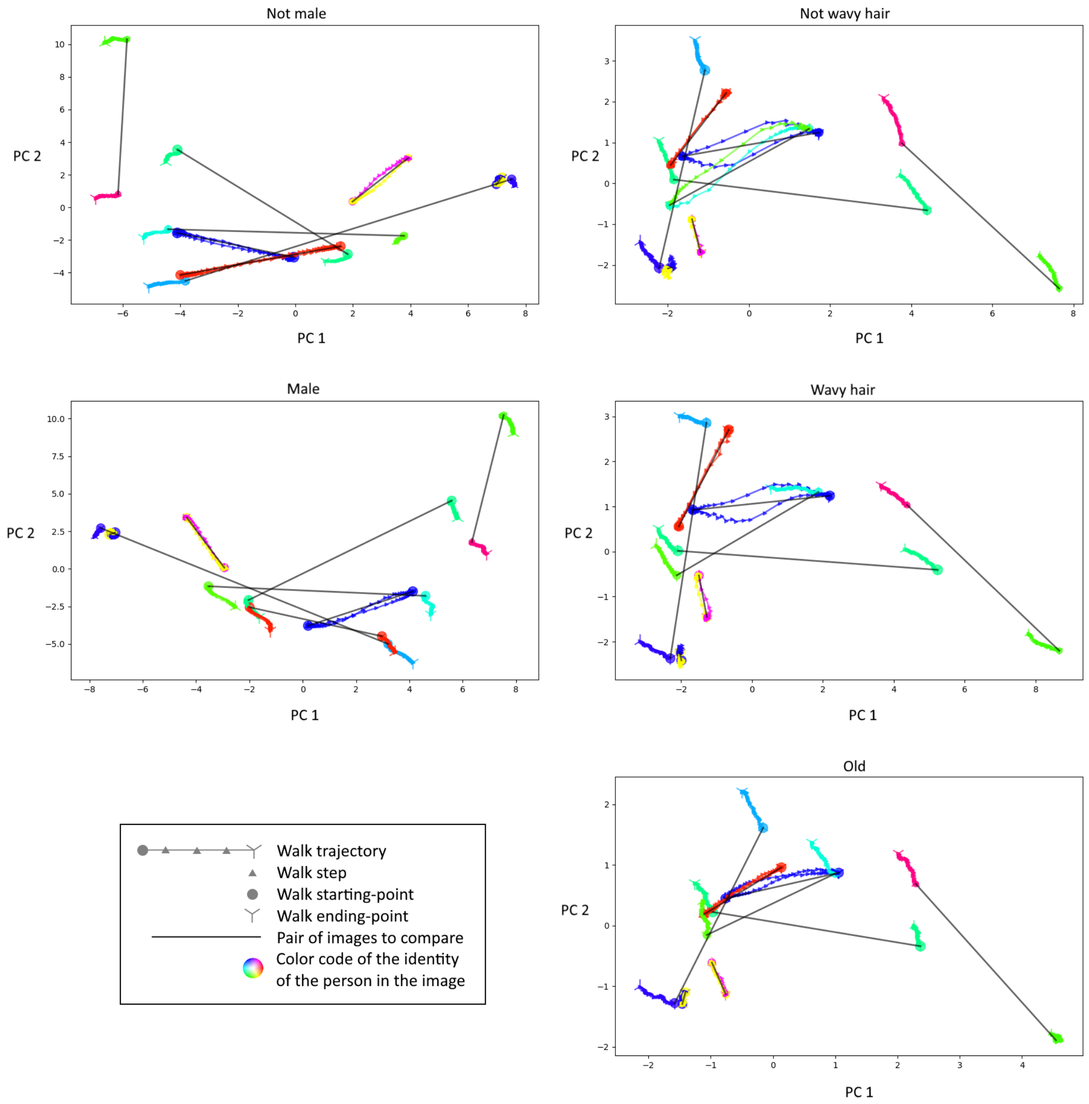

Figure 8. Embedded space walks on 5 model configurations.

Parameter Tuning

On the plots of the embedded space walks we can see when too many or too few pairs reach each other, and if the walks move too slowly or too fast. In the first case one would tune the parameters that define the valley of attraction β and σ, in the second case the walking step size α can be tuned. In case of Figure 8 one can discover that the probability that a pair connects is dependent on the distance between the starting-points, this indicates that the attraction is too strong and more weight should be given to the energy landscape, which is achievable by increasing the walk step size α and decreasing the depth of the valley of attraction β or the width of the valley of attraction σ. A unique set of parameters was needed for every model configuration to achieve the best parametrization.

In Figure 9 we can see the walks on the models for not male, old and not wavy hair images with tuned parameters. The accuracy with good parameters on this data is 75%. One of pairs with identities 2 and 3 has reachability in one direction on the not male and not wavy hair configurations, but it will be classified correctly since the maximum of the two directions is returned.

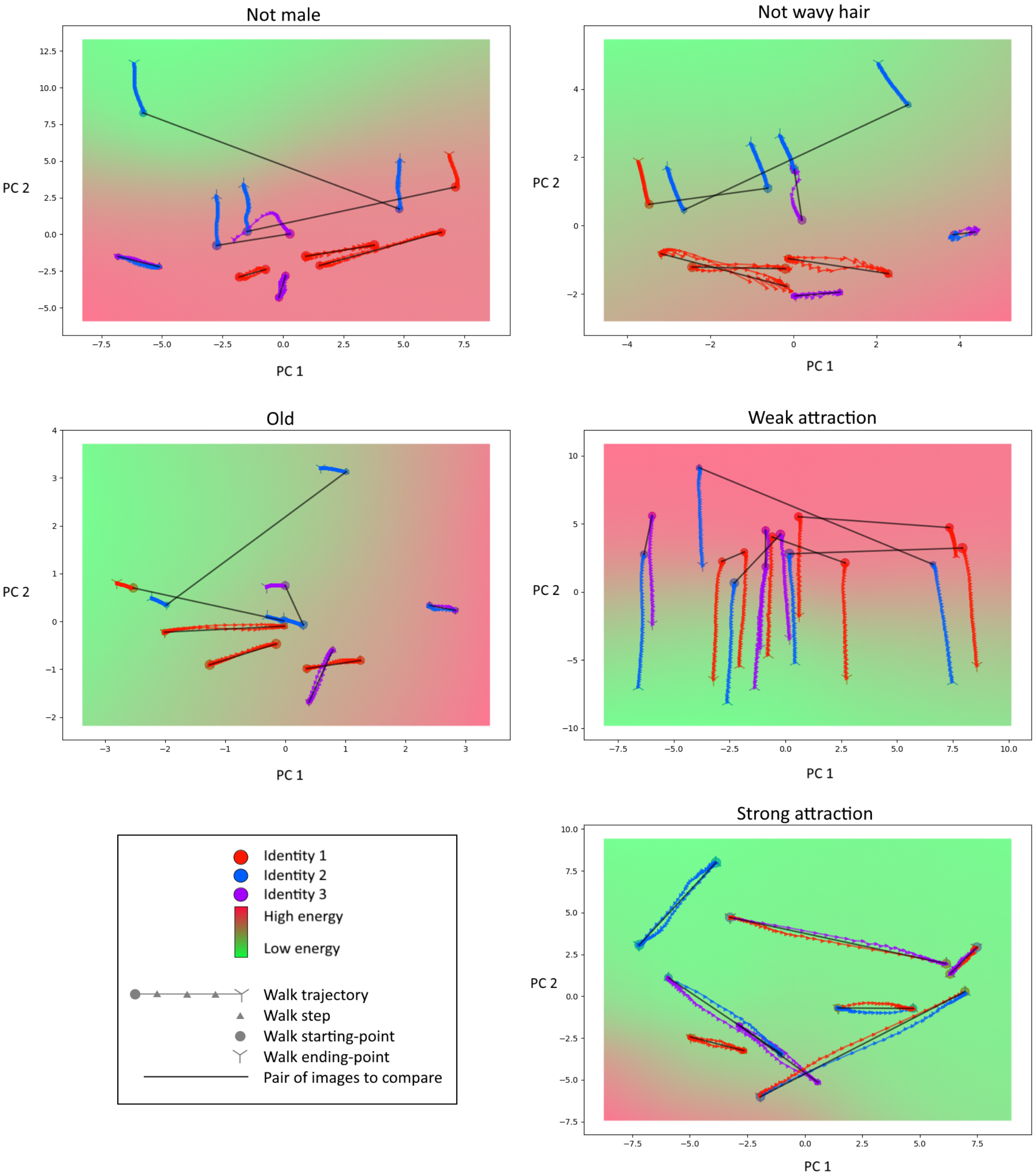

Figure 9. Method performance with tuned parameters on the models of not male, old and not wavy hair images, and the effect of using too strong or too weak valley of attraction.

In Figure 9 we can also see the effect of low values of β and σ on the subplot of weak attraction. We observe that when β and σ are low, the walks will be directed in a direction that decreases the energy regardless of the location of the walk target point. On the subplot of strong attraction one can see that if the opposite is the case, all pairs will always reach each other which is not wanted.

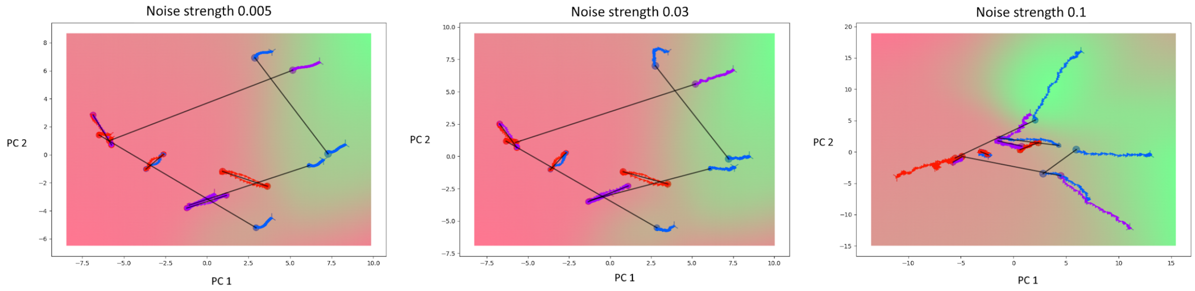

Figure 10. Effect of noise. See the legend of Figure 9 for the meaning of markers.

In Figure 10 we study the effect of the parameter of noise strength on the model for female images. With low and medium noise the method performs with no significant difference. In case of high noise, the walks become visually more unstable and on this test set the reachability within one negative pair of the identities 2 and 3, increased.

Visual Approximations of the Embedded Space Walk Steps

The visual approximations of the walk steps on 3 model configurations are included in the appendix. The visual approximations are images $x_{k,i}$ that have an embedded space representation *Body*($x_{k,i}$) close to $Z_{k,i}$ in terms of euclidean distance, where $k \in \{A, B\}$. On the plots, the walk direction from left to right represents k=A and from right to left represents k=B. In an analogy between the human brain and this method, the visual approximations might correspond to the conscious experiences or imaginations of the brain while thinking if the given face belongs to another person the brain knows.

The Figure 15 of Appendix II depicts the walk steps in Figure 7. According to the plot, the ESV *b* appears to capture the background and texture of the image well as well as the outline of the person, however the details about the facial features like position and shape of the mouth, nose and eyes become heavily distorted and compressed. The probable cause for this is that ESV *b* utilizes only the most downscaled model which in the original model had the purpose of making the overall look of the image consistent. Still, the ESV *b* gave the best results. In Figure 15 we can observe that the visual approximations of the pairs with sample IDs 2, 3, 6 and 8, which according to Figure 7, are connected by the walk trajectory, have a smooth transformation from the image X_A to an image similar to X_B and vice versa.

Figure 16 of Appendix III shows the walk steps of the female model with the ESV *c*. This plot shows that the ESV *c* appears to capture well the facial features the ESV *b* could not, however the overall image composition and background is very different from real images. Both ESV, b and c, use the 4x downscaled submodel but in ESV c the 4x downscaled submodel provides only 128 dimensions to the embedded space whereas in ESV b it provides 1024 dimensions. The use of more abstract features from the submodels could be the reason why the overall composition of the images is not consistent with real images. Another observation from Figure 16 is that all visual approximations look feminine since the underlying model gives low energy for female faces. The walks that were approximated in Figure 16 behaved similarly to the walks in Figure 11.

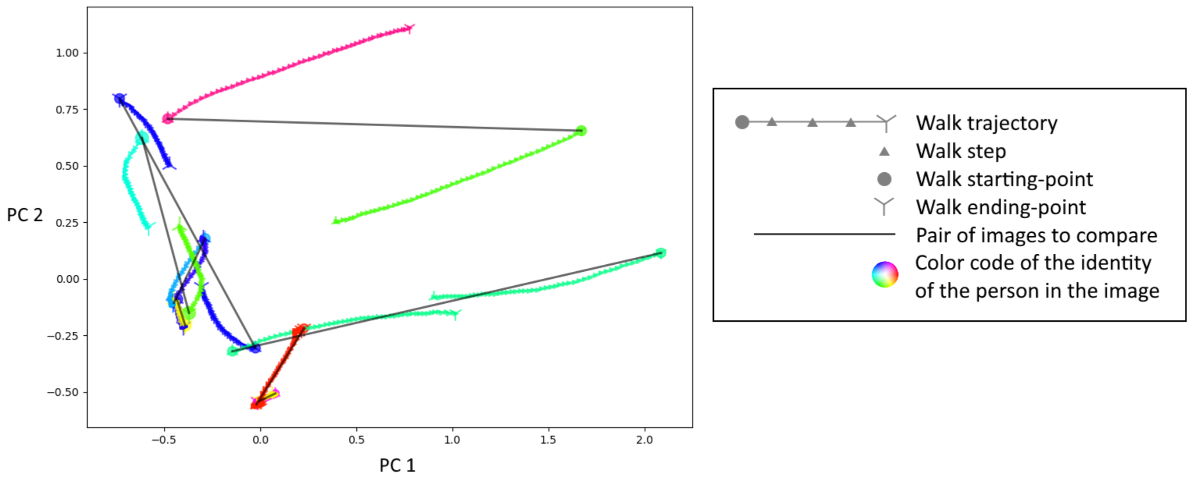

Figure 11. Embedded Space Walks on the Model of Males with Embedded Space Variation c.

In Figure 11 we can see that all walks are directed in the right direction, however some avoid a direct approach to the target. Strangely, the walk trajectories within pairs are much more symmetrical than in the case of using ESV b. Further experimentation showed that . The Figure 17 of Appendix IV depicts the walk steps in Figure 11. Similarly to Figure 16, Figure 17 plot shows that ESV c is good at capturing facial details but not texture and background. All visual approximations in Figure 17 look masculine since the model for males was used during the walk.

4.2. Objective Assessment on Method Performance

Since the method is still experimental and has not been perfected, the visual analysis will be more meaningful than measuring the accuracies and metrics of the model configurations. Nevertheless, we present some formal metrics to make the performance of the method better comparable to other methods for the same task.

In Figure 12 we give the accuracy and F1-score of predictions on the pair distances before applying the method so that we will know how much these baseline results improve by applying the method. These results also show that by transforming the input to the embedded space, the data is already weakly separable.

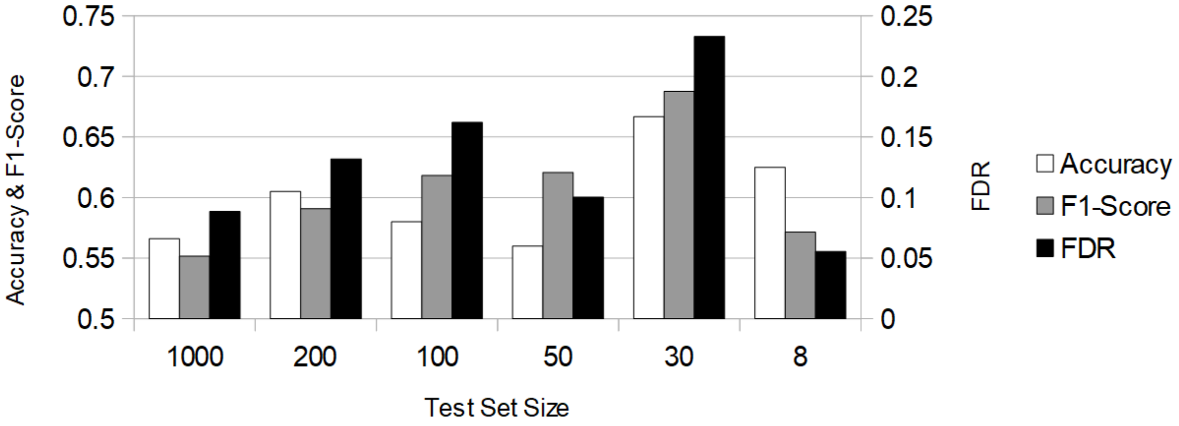

Figure 12. Baseline Metrics without Applying the Method on Various Test Set Sizes on the Embedded Space Variation *b*.

The model configurations which used the ESV b were most successful. In Figure 13 there are performance metrics of the 4 best model configurations with the ESV b. The test set sizes of the model configurations 1, 2, 3, 4 are respectively 50, 50, 50, 30. The configuration 4 seems to be the best with accuracy of 0.7, however the baseline performance for the test set size 30 is also high. Although, configuration 4 has still the highest rating value of all experiments that have been run, indicating that the method with the model configuration 4 improves the separability of the data. See Table 1 of Appendix V for full details on the model configurations and performances.

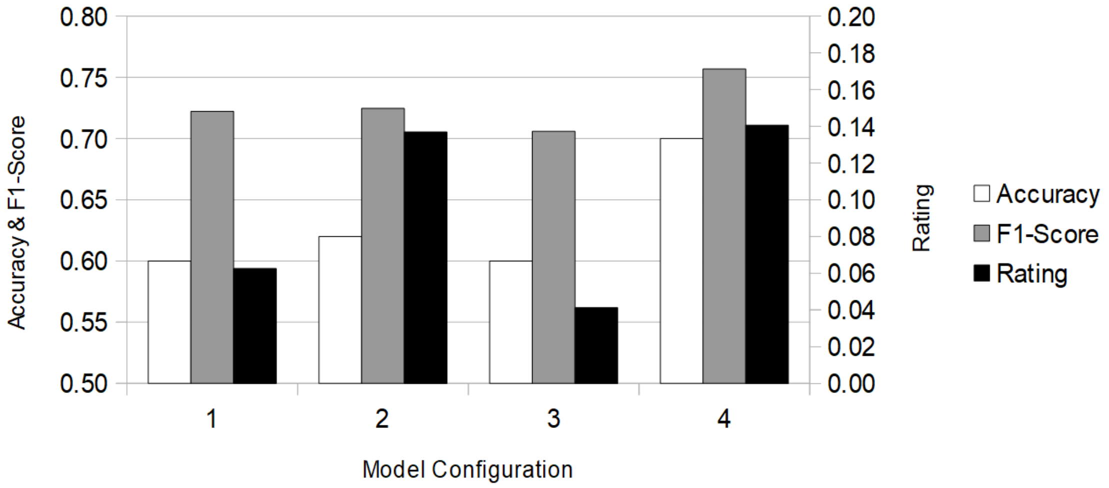

Figure 13. Performance of Models with Embedded Space Variation *b* Compared via Rating.

The ESV *c* failed to give positive ratings on any model configuration that was tried. The ESV *a* gave small positive ratings in some cases, which are shown in Figure 14. Interestingly, the ESV *a* resulted in an exceptionally high FDR ratio in one case, where the value of FDR_{pre} was very low and a small increase in FDR was sufficient to result in high FDR ratio. The test set sizes of the ESVs and model configurations *b*, *b*, *a*, *a*, *b* in Figure 14 are 50, 50, 100, 100, 30 in the same order as on the plot. The accuracy and F1-score of ESV *a* is still lower than that of ESV *b*. The poor performance of ESVs *a* and *c* might be caused by the fact that their *Top* submodel is a linear transformation.

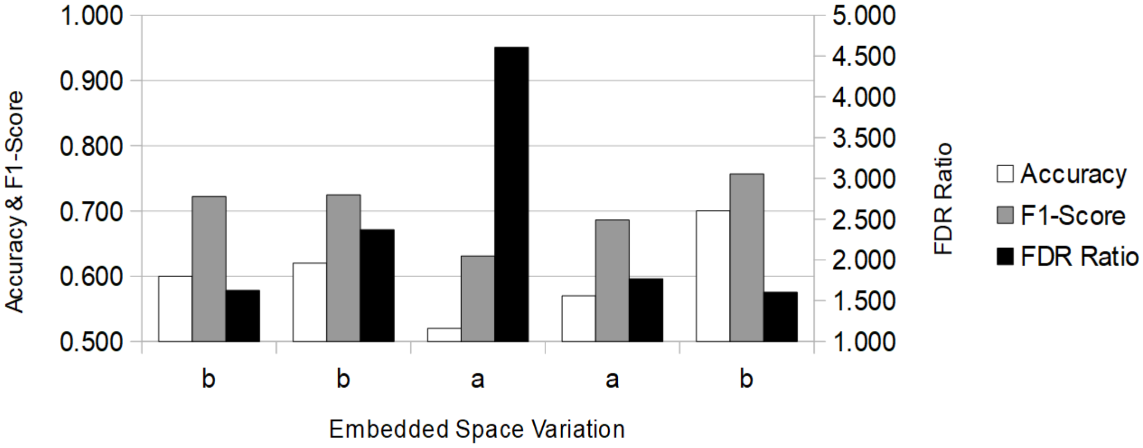

Figure 14. Performance of Models with Embedded Space Variations *a* and *b* Compared via FDR Ratio.

These results are explicitly disclosed in Table 1 with parameters to reproduce them, the test set sizes and the additional measure of post-walk FDR. Table 1 is composed of the results that have a top Rating or a top FDR Ratio. From the table we can observe that the ESV *a* has a much lower FDR, which explains the accuracies at the baseline level.

5. Discussion

Our results demonstrate that, firstly, in the embedded space of the EBM, pairs of common identity are on average placed closer than pairs of different identities. Secondly, by using PCA to visualize the walk trajectories, it is possible to manually intelligently tune the model parameters, as opposed to grid or random search. If this process could be automated, it would make the study of this method more efficient and potentially lead to better results. Thirdly, it is possible to visually approximate the embedded space vectors in the input space if the image of a nearby embedded space vector is known. This can be applied sequentially to visually approximate longer walk trajectories. Fourthly, it is possible to improve the separability measure (FDR) of the positive and negative classes on euclidean distance 1.5-2.3 times by using this method.

The models we used in this experiment were trained to estimate the probability that the input belongs to a part of the dataset. It is interesting that an increase in the class separability is achievable via the use of the output of such models on a different task. Though the models were intentionally selected to not to be designed for this task at hand specifically, a single model for the whole dataset would have been preferred, or alternatively a large number of models for various features like male, old, wavy hair and so on, which could be combined to output an energy landscape, better suited for the task. As a possible improvement, disjunctive composition of the EBMs like in [20], could be used to better combine the models, or a new EBM could be trained on the whole dataset for this specific purpose, removing current limitations on the ways to define embedded space. In future work, it could also be interesting to compare the Langevin steps during the image generation process [16] to the walk trajectories between the pairs in the embedded space.

If the FaceNet or similar model were to be trained with an energy-based head network for image generation built on the hidden space, then the use of our proposed algorithm to navigate the hidden space, might lead to even more accurate results. If improved, the proposed method could be applied on other tasks besides face recognition as well by first training an EBM on data and afterwards defining some task as a navigation problem on the energy landscape.

6. Conclusion

In this thesis we have proposed an algorithm to perform recognition via local navigation in a hidden vector space of an image processing neural network. Further, a baseline for the performance of the algorithm was determined and the algorithm was evaluated. Additionally, steps of the loops in the algorithm were visualized to gain more insight on the capabilities of the method. Therefore, the aim of the thesis – to implement, evaluate and analyse a method to solve face recognition as a problem of navigation – was achieved. We think it would be interesting to probe and improve this method further, exploring better model architectures and compositions, parameter tuning automation and characteristics of various embedded space variations.

7. References

[1] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, ‘Building Machines That Learn and Think Like People’, *ArXiv160400289 Cs Stat*, Nov. 2016, Accessed: May 03, 2021. [Online]. Available: http://arxiv.org/abs/1604.00289.

[2] F. H. Sinz, X. Pitkow, J. Reimer, M. Bethge, and A. S. Tolias, ‘Engineering a Less Artificial Intelligence’, *Neuron*, vol. 103, no. 6, pp. 967–979, Sep. 2019, doi: 10.1016/j.neuron.2019.08.034.

[3] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, ‘How Does the Brain Solve Visual Object Recognition?’, *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012, doi: 10.1016/j.neuron.2012.01.010.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, ‘FaceNet: A unified embedding for face recognition and clustering’, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[5] Y. Du, S. Li, J. Tenenbaum, and I. Mordatch, ‘Improved Contrastive Divergence Training of Energy Based Models’, *ArXiv201201316 Cs*, Apr. 2021, Accessed: Apr. 30, 2021. [Online]. Available: http://arxiv.org/abs/2012.01316.

[6] A. Yuille and D. Kersten, ‘Vision as Bayesian inference: analysis by synthesis?’, *Trends Cogn. Sci.*, vol. 10, no. 7, pp. 301–308, Jul. 2006, doi: 10.1016/j.tics.2006.05.002.

[7] J. Wäldchen and P. Mäder, ‘Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review’, *Arch. Comput. Methods Eng.*, vol. 25, no. 2, pp. 507–543, Apr. 2018, doi: 10.1007/s11831-016-9206-z.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, ‘DeepFace: Closing the Gap to Human-Level Performance in Face Verification’, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.

[9] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, ‘Deep convolutional networks do not classify based on global object shape’, *PLOS Comput. Biol.*, vol. 14, no. 12, p. e1006613, Dec. 2018, doi: 10.1371/journal.pcbi.1006613.

[10] D. R. Hofstadter and E. Sander, *Surfaces and essences: analogy as the fuel and fire of thinking*. New York: Basic Books, 2013.

[11] J. L. S. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller, ‘Navigating cognition: Spatial codes for human thinking’, *Science*, vol. 362, no. 6415, p. eaat6766, Nov. 2018, doi: 10.1126/science.aat6766.

[12] T. E. J. Behrens *et al.*, ‘What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior’, *Neuron*, vol. 100, no. 2, pp. 490–509, Oct. 2018, doi: 10.1016/j.neuron.2018.10.002.

[13] D. Marr, *Vision: a computational investigation into the human representation and processing of visual information*. Cambridge, Mass: MIT Press, 2010.

[14] A. Jahanian, L. Chai, and P. Isola, ‘On the “steerability” of generative adversarial networks’, *ArXiv190707171 Cs*, Feb. 2020, Accessed: May 01, 2021. [Online]. Available: http://arxiv.org/abs/1907.07171.

[15] J. Hopfield, ‘Neural Networks and Physical Systems with Emergent Collective Computational Abilities’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, pp. 2554–8, 1982, doi: 10.1073/pnas.79.8.2554.

[16] Y. Du and I. Mordatch, ‘Implicit Generation and Generalization in Energy-Based Models’, *ArXiv190308689 Cs Stat*, Jun. 2020, Accessed: Apr. 30, 2021. [Online]. Available: http://arxiv.org/abs/1903.08689.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang, ‘Deep Learning Face Attributes in the Wild’. Dec. 2015, Accessed: May 02, 2021. [Online]. Available: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[18] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, ‘A Tutorial on Energy-Based Learning’, p. 59.

[19] ‘Fisher Linear Discriminant - an overview | ScienceDirect Topics’. https://www.sciencedirect.com/topics/engineering/fisher-linear-discriminant (accessed May 06, 2021).

[20] Y. Du, S. Li, and I. Mordatch, ‘Compositional Visual Generation and Inference with Energy Based Models’, *ArXiv200406030 Cs Stat*, Dec. 2020, Accessed: Apr. 30, 2021. [Online].

Available: http://arxiv.org/abs/2004.06030.

Appendix

I. License

Non-exclusive licence to reproduce thesis and make thesis public

I, Henri Harri Laiho,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Recognition as Navigation in Energy-Based Models,

(*title of thesis*)

supervised by Raul Vicente Zafra, Jaan Aru, Tarun Khajuria.

(*supervisor's name*)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Henri Harri Laiho

06/05/2021

II. Visual Approximations of Walks

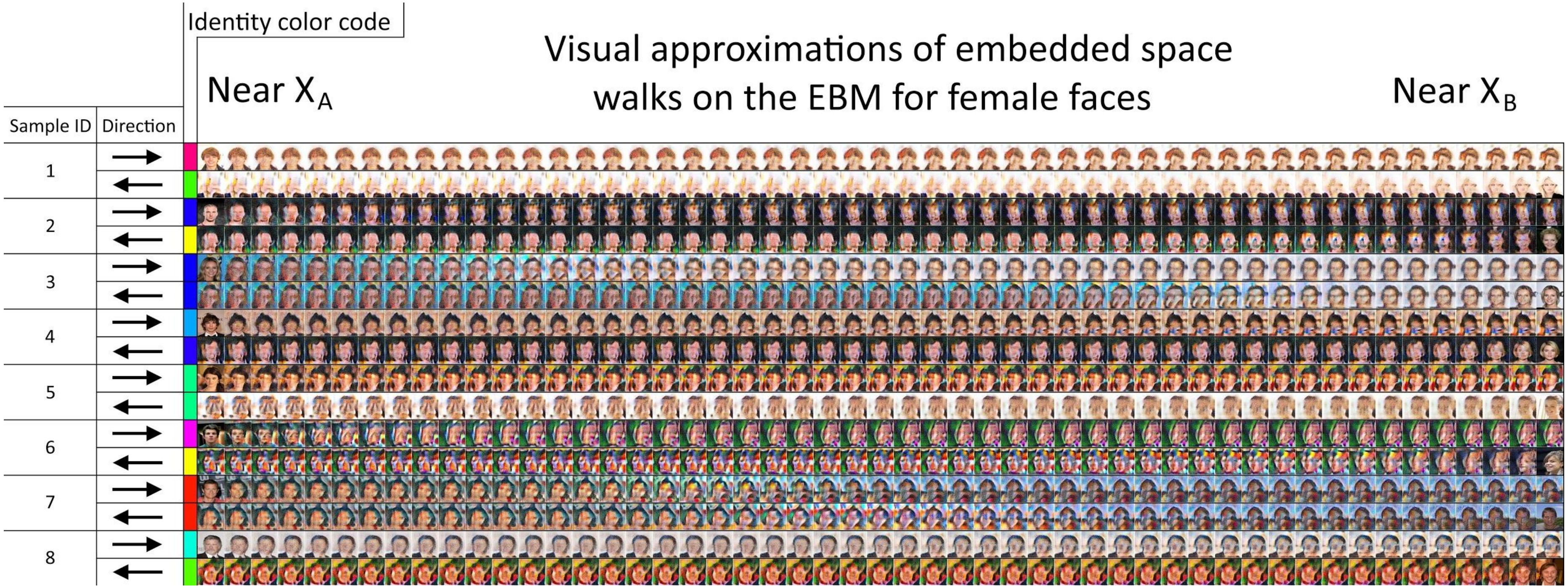
Identity color code
Visual approximations of embedded space walks on the EBM for female faces
Near X_A
Near X_B
Sample ID
Direction
1
2
3
4
5
6
7
8

Figure 15. Visual Approximations of the Walk Steps on the Model for Females using Embedded Space Variation b

III. Visual Approximations of Walks

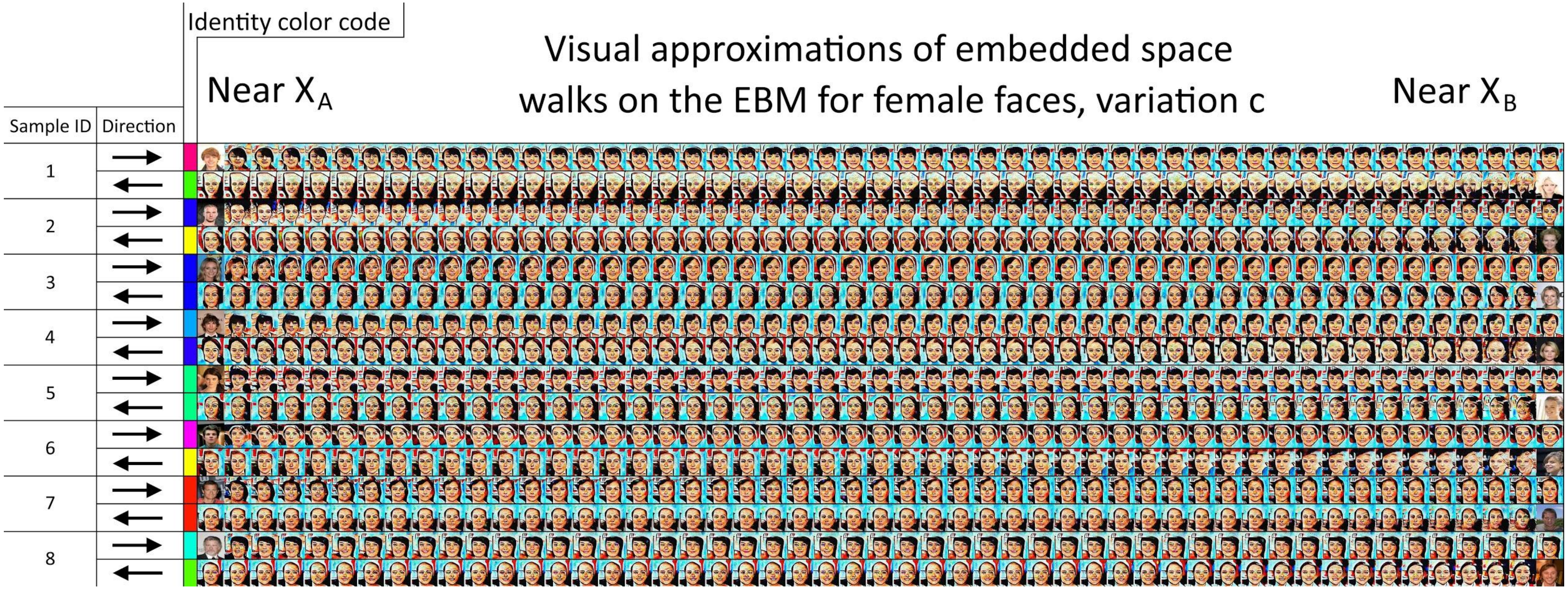
Identity color code
Visual approximations of embedded space walks on the EBM for female faces, variation c
Near X_A
Near X_B
Sample ID
Direction
1
2
3
4
5
6
7
8

Figure 16. Visual Approximations of the Walk Steps on the Model for Females using Embedded Space Variation c

IV. Visual Approximations of Walks

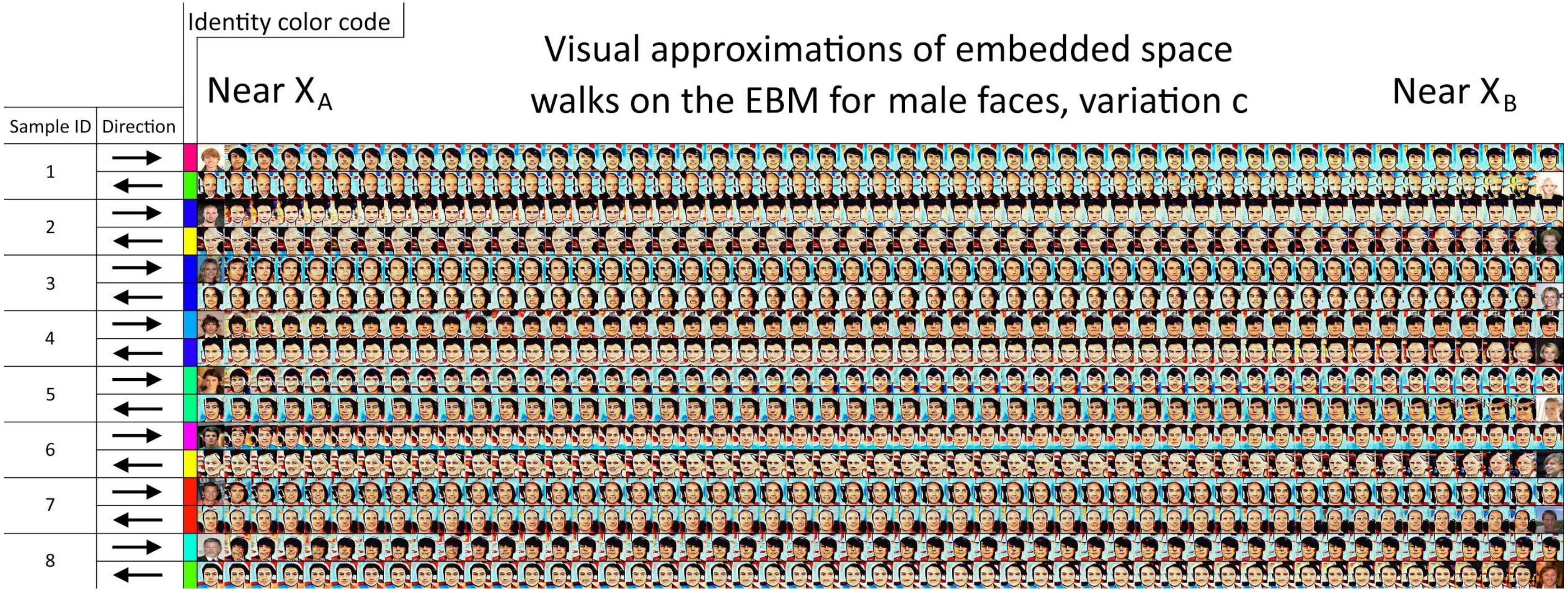
Identity color code
Visual approximations of embedded space walks on the EBM for male faces, variation c
Near X_A
Near X_B
Sample ID
Direction
1
→
←
2
→
←
3
→
←
4
→
←
5
→
←
6
→
←
7
→
←
8
→
←

Figure 17. Visual Approximations of the Walk Steps on the Model for Males using Embedded Space Variation c

V. Parameters for the Best Results

Table 1. Best Model Configurations and Their Metrics

ESV	Models	α	β	σ	Noise Strength	Test Size	Accuracy	F1-Score	FDR	Rating	FDR Ratio
b	Female	50	7	34	0.001	50	0.6	0.722	0.163	0.063	1.626
	Male	50	7	34							
	Old	50	7	34							
	Wavy Hair	50	7	34							
b	Female	50	7	34	0.001	50	0.62	0.725	0.237	0.137	2.369
	Male	50	7	34							
	Old	50	7	34							
	Wavy Hair	50	7	34							
a	Female	0.1	0.7	0.25	0.01	100	0.52	0.631	0.002	0.0014	4.607
	Male	0.1	0.7	0.25							
	Wavy Hair	0.05	2.5	1							
	Not Wavy Hair	0.05	2.5	1							
a	Female	0.1	0.35	0.13	0.01	100	0.57	0.686	0.001	0.0003	1.768
	Male	0.1	0.35	0.13							
	Wavy Hair	0.05	1.25	0.5							
	Not Wavy Hair	0.05	1.25	0.5							
b	Female	5.9	22	12	0.01	30	0.7	0.757	0.373	0.141	1.604
	Male	5.9	16	16							
	Old	1.2	1.6	1.6							
	Wavy Hair	4.8	2.8	2.8							
	Not Wavy Hair	4.8	2.7	2.7							
b	Male	5	4	5.5	0.001	50	0.6	0.706	0.122	0.041	1.507
	Wavy Hair	5	4	5.5							