PID 53 - Obfuscate: One-shot Face Verification

Micha de Groot (10434410)[1] **Pieter Kronemeijer** (11064838)[1] **Anton Steenvoorden** (11850493)[1]
Douwe van der Wal (11042206)[1] **Laurens Weitkamp** (11011629)[1]

Abstract

In this paper, we propose a novel pipeline that combines a face verification model with face augmentation methods for the task of one-shot face verification. The target application is face verification on event images to ensure privacy of visitors. The pipeline is tested for augmentations in pose, illumination and expression using the Multi-PIE and RaFD datasets. We show that our method can increase recall significantly, but that this is dependent on the quality of augmentations. We furthermore provide insight as to where and why the pipeline might fail.

1. Introduction

In the current day and age privacy is a much discussed topic. With the new European GDPR law (Council of European Union, 2016) a new constraint in the world of public events was introduced. Photographs of individuals can no longer be published without explicit approval. Organizers can either demand approval before allowing the visitor into the event, or come up with some method to remove people who request anonymity from the picture before it can be posted.

A current solution involves having visitors wear a sticker on their forehead[1], and checking for this mark when placing pictures online. Having to wear a sticker to get privacy is of course far from an ideal solution, as the sticker can fall off, has to be detected by people and requires effort for privacy which should be the norm. Currently, no user friendly system is available which automates this for an event.

To address this problem, we propose a visitor-friendly approach for facial verification based on a single neutral frontal image of a face, a portrait, taken at the start of an event in a controlled environment. This task is also known as one-shot verification (Fei-Fei et al., 2006). The image will be augmented with various illuminations, viewpoints, expressions and glasses to maximise recall in images taken during the event. Only anonymous face embeddings will be stored, as to preserve privacy. By using augmentations instead of real images, we can keep the process of adding someone to the database quick and simple while still having sufficient samples of a person to recognise them.

To achieve this, we will build upon the work in (Miraftabzadeh et al., 2017), which uses embedding models for the task of face verification in crowds. We extend this work by adding augmentations of facial images to the embedding space through deep learning models in order to get better performance compared to having only a single image of the target. Data augmentation for facial features has been discussed in (Lv et al., 2017; Masi et al., 2016), which proposed adding a number of facial augmentations to reduce training time for convolutional neural networks. Our work is similar, but we instead use augmentations specifically to reduce the number of images needed for the task of face verification in embedding models. Our work uses a state-of-the-art face-verification embedding model called FaceNet (Schroff et al., 2015) which claims to be both illumination and pose invariant.

We define the problem as follows: using only a single frontal face portrait, how well does the model recognise the same portrait in varying poses, illuminations and and changes in expressions. From this, two questions naturally arise:

1. How robust is the model to changes in pose, expression and illumination?
2. Can we improve the accuracy by appending augmented versions of the portrait to the embedding space (i.e., using machine learning to augment the frontal portrait to allow for different poses, expressions and illuminations)?

As mentioned, FaceNet claims to be both illumination and pose invariant. We investigate this claim by using a single neutral image to match with, whilst varying poses and illuminations are input. We hypothesise that adding augmentations to the embedding space will yield an increase in performance, despite the claimed invariances.

A demo of our work is available at http://pindakaas.ga:8888.

[1] Article on using stickers to recognize people in pictures https://bit.ly/2nNqtTD

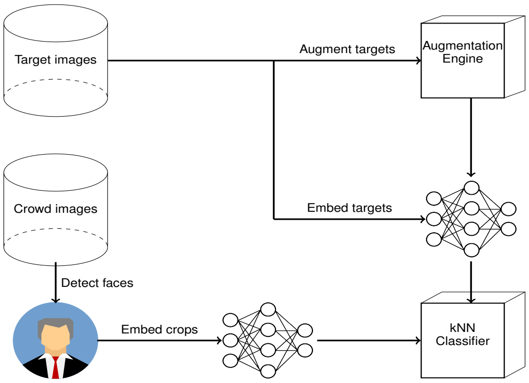

Figure 1. Visualisation of pipeline. For target images, we extract one face, augment this, and encode both into embedding space. Crowd images are processed by locating multiple faces and encoding each face into embedding space. We then perform kNN matching with the target embeddings.

2. Method

2.1. Face Verification Pipeline

The task of identifying people in crowds is split in two parts. First, target images are acquired from which the face is extracted using a face detector. This face is augmented with various poses, illuminations, expressions and glasses, and added to the target dataset. From the crowd image all faces are detected and processed, but not augmented. Second, the face verification is performed by embedding the crowd and target images using the embedding model. The matching is then performed in Euclidean space with the embeddings of the crowd images compared to the targets and their augmentations. A schematic visualisation can be seen in Figure 1.

2.2. FaceNet

We will use FaceNet to encode faces into an embedding space. FaceNet is trained to be invariant for poses and illumination changes, and learns to efficiently embed face images in a high-dimensional space where distance directly corresponds to face similarity. Our results, displayed in Figure 2, show that FaceNet is not fully pose and illumination invariant and that there is room for improvement. Invariance is achieved by using triplets of training samples, which consist of a random sample (anchor), another sample of the same person (positive sample), and a random sample of a different person (negative sample). FaceNet employs a loss function that forces the embedding of the positive example to have a small distance to the random sample embedding, and a large distance with respect to the negative example. The FaceNet model used in this paper has been trained on the VGG Faces dataset (Parkhi et al., 2015). Both detection and preprocessing of faces in crowds is done using a Multi-task Cascaded Convolutional Network (MTCNN) (Zhang et al., 2016) (For more details about preprocessing and FaceNet we refer the reader to Appendix A).

2.3. Evaluation

We evaluate our methods on the datasets described in Section 3.1. The datasets are split into two groups we call ”source” and ”target”. Images in *source* are unseen, while *targets* are images we want to match with. The model randomly selects n target actors from the total number of actors. In the Multi-PIE dataset 100 actors are selected. In the RaFD dataset 50 actors are selected. Afterwards only the neutral image setup is used as baseline image for each actor. The number of target images is increased when using variations or augmentations, due to the smaller size of RaFD when compared to Multi-PIE. The images added to the targets are kept out from the source set. This results in one source set for the baseline, variation and augmentation setup.

All source images are matched to their nearest neighbour. If the distance to their nearest neighbour is above the predetermined threshold it it classified as not having a target.

To measure performance, accuracy and recall are used. We define a positive example as an input image that has a corresponding target image. A negative example is an input image that has no corresponding target.

In general the total classification accuracy is the measure used in identification tasks. However, the recall measure is of higher importance in this task, as we do not want to miss any people that want to be blurred out from a photo in a hypothetical event. Our task is thus to maximise recall while maintaining a high accuracy.

2.4. Privacy, Security, Data Governance, Algorithm Reliability

Privacy is important and we have been made aware of this by several notable events, e.g. Facebook has shared data of many of its users without informing the users[2]. As the main focus of this project is ensuring privacy to public event goers, the technology used to provide this service should treat the data used very carefully. Our method does not require to store the images taken at the event, as we can simply forward it through our network and save the resulting embeddings only. The photographer is in charge of securing his copy of the crowd photo with the visible faces. Our model ensures the image used to blur the face is absolutely private. Moreover, the pipeline of this project is optimised

[2] Article on Facebook’s privacy infringement https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html

for recall, meaning that the main goal is to blur everybody that wanted to be blurred, while caring less about the accidental blurring of people that did not necessarily want to be blurred. The algorithm consistently scores around 80% accuracy on average, which means that people might need to manually verify tagged images.

3. Experiments

3.1. Datasets

To measure the effects of augmented poses, illuminations and expressions we need a dataset that contains multiple photos of people, taken in different poses, with different illuminations, and with different facial expressions. Two datasets satisfy the description partially: Multi-PIE (Gross et al., 2010) and RaFD (Langner et al., 2010).

3.1.1. Multi-PIE

Multi-PIE has variations in person, illumination, viewpoint and session, where every viewpoint of the same illumination and session is taken at the same time. Furthermore, this dataset contains three different expressions, but these were not used as another dataset was more suitable for expression testing

We used a subset of this dataset with viewpoint variations between -60 and 60 degrees horizontally in steps of 15 degrees, and 0 degrees vertically (only eye-height). Illuminations only include left, frontal, right, and full illumination. All 337 people, (max 4) sessions were used.

Additionally, in this dataset some people wear glasses, but have taken them off or wear a different pair in other sessions. Whether glasses are worn was not annotated, so people that took off or put on glasses in between sessions were separated by hand for a specific test case.

3.1.2. RaFD

The second dataset is The Radboud Faces Database (RaFD), which contains horizontal variations in viewpoint from -90 to 90 degrees in steps of 45 degrees. It has 67 models with 8 different emotions. Only frontal gazes were used during this project. We have specifically added RaFD in addition to Multi-PIE because the number of facial expressions in Multi-PIE is relatively low (only three types).

3.2. Baselines

The baseline experiment splits the data and does not add any images to the singular targets. The target images are thus only the neutral images, as defined earlier. The input images are all remaining images.

3.3. Improving Recall

Multiple experiments are performed to investigate performance of the model. These experiments are split in three parts. First, a baseline is obtained. Second, to see whether improvements can be made, the target dataset is extended with true variations of the selected actors. Finally, the effect on performance when extending the target dataset with artificial variations, we call augmentations, is investigated. The variations and augmentations investigate the effect of adding pose changes, illumination changes and expression changes. The experiments are performed with the same sources in to allow comparison. A line search over threshold values for the matching is performed to find the best performance, the best performing value will be used in further comparisons.

3.3.1. Pose augmentations

The pose augmentations are made using a textured reconstruction of the face in the image which is then rotated whilst keeping the background in tact. The augmentations made are, for the yaw, -45, -30, -15, 15, 30 and 45 degree rotations.

3.3.2. Lighting augmentations

The lightning augmentations are made using a textured reconstruction of the face in the image, after which a new light source is added. The background is lost in this process. The intensity of the light source is augmented for several levels. The augmentations are performed only on the neutral faces. The augmentations are light from the left, right and front, with intensities: 1, 3.5, 6 Watt/m^2.

3.3.3. Expression augmentations

The expression augmentations are made using the GANimation (Pumarola et al., 2018) network. GANimation can generate anatomically-aware facial expression changes from a single image, conditioned on another image used to extract the expression from.

3.3.4. Face Occlusions through glasses

During classification we discovered that the models have a lower accuracy when the target people are wearing glasses, as the top 10 errors in the baseline runs contain almost exclusively faces with glasses. A hypothesis is that the error is not specifically tied to glasses, but rather face occlusion in general. To test this, we will investigate the distance between both embeddings of augmentations of faces with occlusions and actual glasses, with faces that have no glasses. Occlusions are made by detecting facial landmarks with dlib (King, 2009), and placing a black rectangle over the eyes, see Figure 5 for some examples.

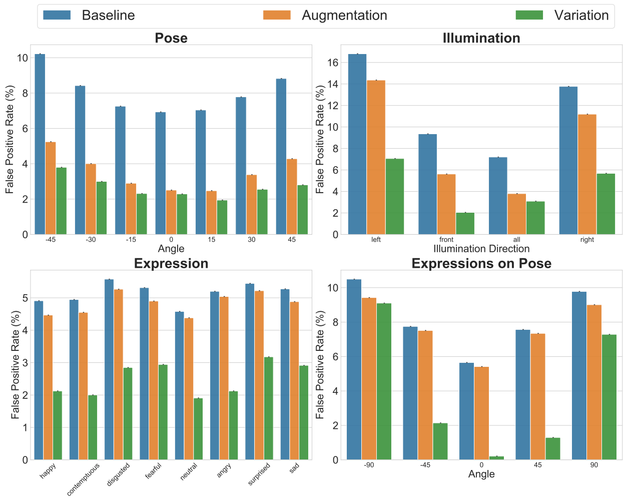

Figure 2. Number of false negative misclassifications made in labels pose, illumination, expressions and expressions on pose, lower bars indicate an increase in recall. For each label, an increase in recall can be seen when adding either variations or augmentations. Expressions on pose indicates the impact of adding *expression* variations or augmentations to the targets for the various *poses*. It is notable that the errors for pose angle -15 becomes nearly 0.

3.3.5. Classification breakdown

To analyse the exact impact of augmentations and their comparison with the variation images the recall score for each test is split up into the different modes of each image. That way we can see if augmenting an image with a different pose actually results in correctly classifying more images with a different pose.

4. Results

Various experiments have shown that the best increase in recall through augmented illumination and poses can be obtained by adding as much variations or augmentations that are available. For the pose and illumination this means adding all values listed in section 3.3.1 and 3.3.2 respectively. The improvement in recall for the expression augmentations levels off when two different expressions have been added, which can be explained by the fact that the different expressions are independently generated. The exact improve can be seen in Table 1 and Figure 2. Here we see that both the pose and illumination recall error show an upward curve towards the images that vary more from the neutral image. The data shows that the augmentations and variations show a similar curve, though it is more gentle. For the expressions there is no such curve because expressions can't be said to vary from a neutral expression in a linear fashion.

4.1. Robustness of the model

In Section 2.2 it was mentioned that FaceNet claims to be viewpoint and illumination invariant. In the top row of Figure 2 it is visible that the baseline runs for frontal viewpoint and full illumination have less errors than the different viewpoints and illuminations. Clearly, FaceNet is not illumination and viewpoint invariant, it is robust at best.

As described in Section 3.1, there are people wearing glasses in Multi-PIE, but this is not annotated. By analysing the images of the top 5 people that are most often misclassified, we notice that 4 of them wear glasses on their target image. The only images that are correctly classified, usually have frontal or full lighting, and most other images are not matched to any face. We can see in Figure 2, that the model is not invariant to extreme pose changes, such as angles at -45 and 45 degrees.

4.2. Recall Improvement

The results achieved after adding both variations and augmentations to the target dataset can be seen in Table 1. Recall is improved both when using variations and when using augmentations, but this comes at the cost of precision. Adding the variations can however lead to both an increase in recall *and* accuracy, as we can see in the cases of pose and expressions.

If we look at Figure 2, we can see that recall indeed increases when either adding variations or augmentations to the embedding space. Most notably, for pose, the additional embeddings significantly increase the model's recall on more extreme angles such as -45 and 45.

To give some insight into why precision is being decreased, we have visualised a case of false-positives and a case of true-positives using t-SNE (van der Maaten and Hinton, 2008), which can be seen in Appendix B.

When augmenting expressions in RaFD, we have seen that misclassifications made on frontal portraits of people are almost completely removed. If we instead look at the more extreme angles, misclassifications are likewise reduced but not significantly. An explanation for this is that the errors for extreme angles ($\sim$90 degrees) are likely caused by the fact that most of the face is not clearly visible anymore, and therefore cannot match the (frontal) target image. Adding expression augmentations does not change this fact.

4.2.1. Face occlusion

Figure 3 shows the average distance between embeddings of images with glasses, without glasses, and with a black bar in front of the face. It is clear that the mean average euclidean embedding distance between faces with glasses and faces without glasses is lower than the embedding distance

Table 1. Recall, precision and accuracy for changes in pose, illumination and expression. Pose and illumination statistics have been calculated on Multie-PIE, statistics for expressions has been calculated on RaFD.

	Pose			**Illumination**			**Expression**		
	baseline	variation	augmentation	baseline	variation	augmentation	baseline	variation	augmentation
recall	0.425	**0.806**	0.733	0.456	**0.791**	0.595	0.236	**0.626**	0.253
precision	**0.401**	0.361	0.327	**0.468**	0.412	0.416	**0.893**	0.885	0.738
accuracy	**0.701**	0.608	0.584	**0.702**	0.632	0.656	0.423	**0.671**	0.415

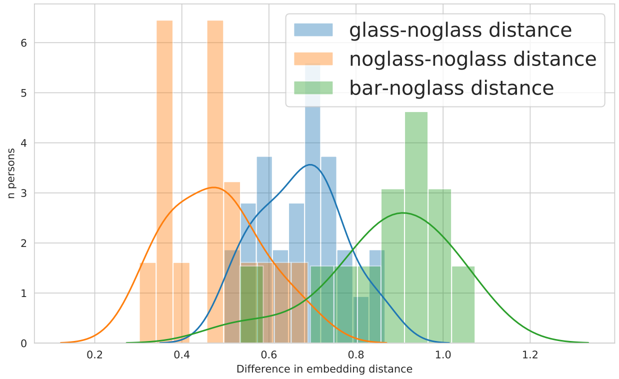

Figure 3. Comparison of average distance between embeddings of faces (same person, different setting) with glasses, without glasses en with a black bar before the eyes

between faces with a black bar and faces without glasses. This means that occlusion by glasses is significantly different from both general occlusion and no occlusion at all, so it might be necessary to augment glasses separately.

5. Discussion

It is theoretically possible to gain a significant increase in recall with a marginal decrease in precision, if it possible that the augmentations can fully capture complex facial features.

With the current augmentations provided by 3DUniversum[3] there is an increase in recall but for some augmentations it is significantly less when contrasted to adding true variations.

The number of false positives however also increases, which could be due to targets being matched to augmentations of different targets. This can be explained further by looking at the embedding space projected using t-SNE, which can be seen in Figure 4 in the Appendix. The augmentations are often not only centred around the embeddings of the original images, but are sometimes located in the area of another target, in Euclidean space.

One approach to increase the recall of the model is finetuning on the dataset. It is possible that the dataset used to train the model with, has some underlying differences in terms of faces, and that finetuning the final layers of the network improves the performance on the dataset we use. We have tried to finetune using just random samples but this failed to increase the performance after a small number of epochs. Finetuning using difficult triplets as training samples also did not increase performance with a small number of epochs. Due to a lack of resources the finetuning has not been further investigated, this is open for future research.

The results of the expression augmentations do not look very realistic, qualitatively speaking, see Appendix D for examples. This might be a reason why the increase in performance when adding augmented expressions is not as much as adding poses.

The experiment with glass face occlusions shows that specific occlusions have their own region in embedding space. Naturally, it would be a good test to see how much improvement glass augmentations give. However, Multi-PIE and RaFD do not provide labels for this; the tests in Section 3.3.4 were run on hand-selected data. Nevertheless, there are datasets that do have labels for glasses, such as CelebA (Liu et al., 2015), and are well suited for this testing.

6. Conclusion

We have performed multiple experiments in which we compare the capabilities of the FaceNet model in for the task of one-shot face verification.

Adding pose, illumination and expression augmentations to the targets increases the recall. If the augmentations are realistically generated (i.e. more like the true variations) a significant increase in recall can be obtained.

The performance can potentially be further enhanced when finetuning is done using augmentations, but this is left for future research.

[3] https://3duniversum.com/

References

Council of European Union (2016). Council regulation (EU) 2016/679 (gdpr). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, 28(5):807–813.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lv, J.-J., Shao, X.-H., Huang, J.-S., Zhou, X.-D., and Zhou, X. (2017). Data augmentation for face recognition. *Neurocomputing*, 230:184–196.

Masi, I., Tran, A. T., Hassner, T., Leksut, J. T., and Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer.

Miraftabzadeh, S. A., Rad, P., Choo, K.-K. R., and Jamshidi, M. (2017). A privacy-aware architecture at the edge for autonomous real-time identity reidentification in crowds. *IEEE Internet of Things Journal*, 5(4):2936–2946.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878.

A. FaceNet Implementation Details

Face detection and preprocessing is done using a Multi-task cascading neural network (MTCNN). The MTCNN splits the work of detecting up into three different stages: (1) calculate a number bounding box proposals and merge large overlapping bounding boxes (known as the P-net), (2) further refining of candidate bounding boxes through pruning (known as the R-net) and (3) calculate 5 landmark points for each candidate bounding box which will centre the resulting candidate (known as the O-net). The resulting size of the image is 160x160x3 pixels, representing width, height and color channels respectively.

B. t-SNE visualisation of augmentations

To give some insight into why the precision might drop whilst the accuracy/recall increases, we have used t-SNE to plot embeddings in two cases in Figure 4.

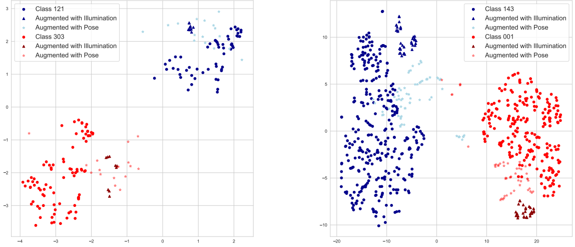

Figure 4. Left: t-SNE embedding of two classes with no false-positive misclassification between either class. Right: t-SNE embedding of two classes with at least one case of false-positive misclassification between either class. In this case, an augmented pose of class 1 was wrongly matched with an augmented pose of class 143.

C. Face occlusion examples

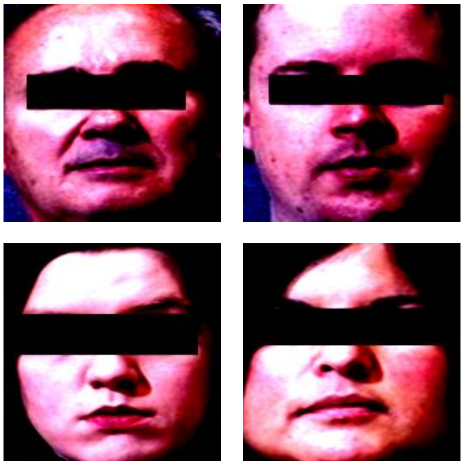

Figure 5. Examples of augmented facial occlusion with a black rectangle in front of the eyes. The bar is places using dlibs facial landmarks

D. Sample of Faces

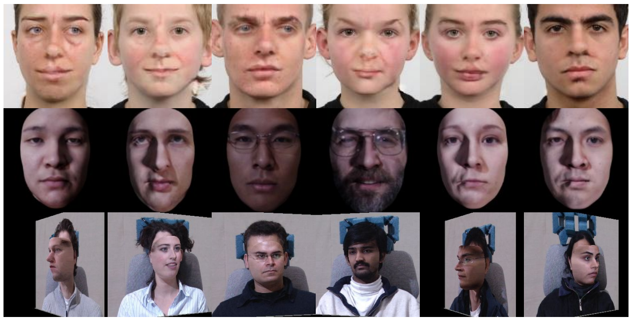

Figure 6. Sample of augmented faces. First row: RaFD faces augmented with expression, second row: Multi-PIE faces augmented with illumination, third row: Multi-PIE faces augmented with pose change.

[1] Supervisors: prof. dr. T. Gevers, dr. S. Karaolgu, PhDs Minh Ngo.

Micha de Groot <micha.degroot@student.uva.nl>
Pieter Kronemeijer <pieter.kronemeijer@student.uva.nl>
Anton Steenvoorden <anton.steenvoorden@student.uva.nl>
Douwe van der Wal <douwe.vanderwal@student.uva.nl>
Laurens Weitkamp <laurens.weitkamp@student.uva.nl>.

Project AI 2019, MSc Artificial Intelligence, Univ. of Amsterdam