Generating IMU data from video using deep neural networks for use in animal activity recognition

JASPER BOVENKERK, University of Twente, The Netherlands

Inertial measurement unit (IMU) data has proven to be quite successful in the field of activity recognition, both on human and animal activities. This IMU data can be gathered relatively easily from animals using collars with sensors. However, for training accurate models a large amount of this data needs to be labeled, which is a very expensive and time-consuming process.

To overcome the issue, researchers have come up with several ways to generate IMU data from video, as labeled video data is abundantly available. Previous approaches mainly make use of pose estimation and forward kinematics. In this paper, however, the viability of using end-to-end deep learning for generating IMU data is evaluated.

This research consists of two parts, the first part will be to use end-to-end deep learning to generate IMU data from video data. The second part will be to train an animal activity recognition(AAR) model to evaluate the effects of adding generated IMU data to the training data of the AAR model.

In this research is shown that, albeit with a fairly small dataset with a limited amount of activities, there are indications that IMU data generated from video using neural networks can contribute to the training of an AAR model.

Additional Key Words and Phrases: Animal activity recognition, inertial measurement unit, video, end-to-end learning, deep learning

1 INTRODUCTION

Animal activity can be a great indicator of many things about the animal and its environment, including the animal's health, environmental events, and social interaction. Collars are already commonly used to track, identify and monitor animals. However, more recent additions to collars are accelerometers and gyroscopes. These allow for new applications in this field, one of which is recognizing animal behavior based on IMU data. [7]

It has already been shown that IMU data can be used to quite accurately recognize animal behavior, like standing, running, and trotting, or human behavior (on the football field in this example), like passing, sprinting, jogging, and shooting [2, 7]. There are still a few limitations though, mainly due to the lack of labeled data to train the model on. The lack of labeled IMU data can be explained by the fact that it is very expensive and time-consuming to label the data.

To remedy this, researchers have looked to generate IMU data from labeled video data, of which significant amounts exist. Commonly used techniques are to extract poses from the video and use forward kinematics to generate IMU data [9, 11, 16]. In this paper the use of end-to-end deep learning will be evaluated for the use in training AAR models.

The question this research aims to answer can be defined as follows:

- How well can IMU data be generated from video using deep learning for the use in training an animal activity recognition model?

Answering this question will consist of 2 parts:

- How well can an end-to-end deep learning model, which was developed using AutoKeras, perform in predicting IMU data based on video?
- How does IMU data generated from video as training data affect the performance of an animal activity recognition model?

For answering these questions, firstly a deep learning model will be developed to generate IMU data from video, and secondly, an AAR model will be made to evaluate to what extent the generated IMU data can contribute to improving the accuracy score of the AAR model.

In section 2 related research will be discussed, including the state of the art in IMU generation and deep learning. In section 3 the dataset and the approach will be discussed. In section 4 the results will be presented and in section 5 conclusions will be drawn from the results.

2 RELATED WORK

Activity recognition has been a field of interest for a long time as it has many uses. Both in the field of human and animal activity recognition. The uses can vary widely and include extracting information on animal health, environmental events and social interaction for animal activity recognition [7] and basing training plans on football match data for human activity recognition [2].

Since it is a popular field of research, many different forms of data have been used for activity recognition. These include, but are not limited to, video [4], sound [14] and IMU data [7, 9].

Of these data types, IMU data is one of the most promising since it can easily be gathered. IMU sensors, such as accelerometers, gyroscopes, and magnetometers, have become small enough to easily fit into wearables, such as phones and watches for humans and all sorts of collars for animals. This is in contrast to video, for example, since it is very hard to have a camera continuously record someone. Video data could also give issues regarding privacy. Also, IMU data has already been successfully used many times to identify activities. Accuracy scores of well in the 90% have been achieved [2, 7].

The classifying algorithms used on IMU data also vary greatly. Random forests [9], naive Bayes [7, 8], decision trees [7] and deep neural networks[11] have all been used in previous research and have been proven capable of activity recognition based on IMU data.

There is a disadvantage to IMU data, however, which is that there is very little labeled data available for training. The amount of labeled data available for other fields, like computer vision and natural language processing, is many times higher than in the field of activity recognition via IMU data[11]. This is unfortunate since the availability of large datasets can improve the models [13].

TScIT 37, July 8, 2022, Enschede, The Netherlands

There are several methods to increase the size of the datasets. First of all, there is the option to just gather more, by manually labeling all data. This is very expensive and time-consuming and therefore not a preferred option. Secondly, there is transfer learning, where a model which was trained for one task with a large amount of data is retrained for another task with limited data. This is hard to apply in this field, however, since it would require some form of pre-existing activity recognition model trained on a large amount of data, but large datasets are hard to come by in this field of research. Lastly, there is data augmentation, which is making new data by slightly modifying existing data or by creating synthetic data from existing data. Both augmentation techniques have already been attempted on IMU data in several ways.

Modifying existing data is a common approach to data in the form of images since scaling, rotating and many other forms of distortions can often easily be applied without changing the meaning of the image. For IMU data, this is significantly more difficult. Nevertheless, this has already been done in some studies [3, 10].

The most common approach for generating synthetic IMU data is to use labeled video data, which is quite abundant. Previous attempts to generate synthetic IMU data from video have delivered promising results[9, 12]. In these papers estimated 2D and 3D positions are used to determine the movement through space and with that the acceleration that takes place.

Previous research also looked into generating IMU data from video for animal activity recognition. In this research, 2D pose estimation was used as well.[16]

Another useful thing to take away from these papers is that they showed that combining the generated IMU data with the actual IMU data allowed them to improve the performance of the animal classification model. Classification based solely on the generated IMU data performed slightly worse than the real IMU data only on the contrary.[9, 11, 16]

The papers mentioned above have shown that pose estimation can be a good way to generate useful IMU data, however, it requires features in the form of specific joints to be selected for use in the IMU generator. The features used here might be suboptimal. To avoid this feature selection end-to-end learning can be used. End-to-end learning is a method where no feature engineering takes place and the raw data is the input for a deep learning model.

There are many different techniques in the field of deep learning, one of interest for this research is Convolutional Neural Network(CNN). CNNs have their origin in image processing problems and were therefore designed based on structures that were observed in the human brain.[2]

CNNs consist of convolutional layers, pooling layers, and fully connected layers, however, the layout and types of these layers can wildly vary. Certain configurations perform well on a specific task, but poorly on another. Therefore it is also important that the right structure is picked. This can be a complex task though.

A solution to this was proposed and made. A new method of neural architecture search(NAS) was proposed and built in the form of the open-source software Auto-Keras. This is a program that will look for the optimal neural network for a specific dataset. [6]

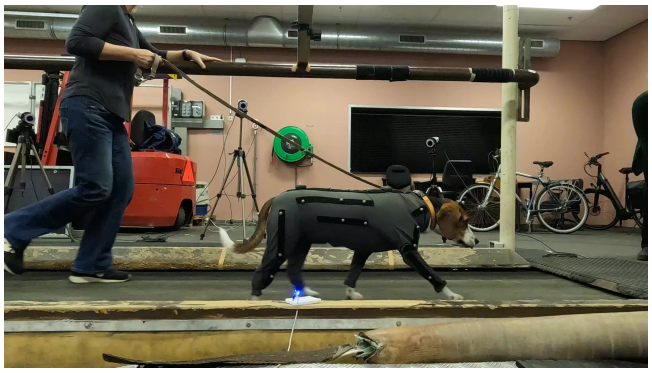

Fig. 1. An example frame from one of the videos taken by the right camera

3 METHODS

3.1 Dataset

The data used in this research was collected by the University of Utrecht and all experiments with the animals complied with Dutch ethics law concerning working with animals. The dataset involves 5 dogs performing 2 activities on a treadmill: walking and trotting. The dogs were of different breeds, but roughly the same size. During these activities, they were monitored by GoPro's from the front, left, right, top and back. These filmed the dogs in 4K at 60/120 Hz (differs per video). In addition, 3 IMU sensors were placed on the dogs, one on the head, one on the withers(shoulders), and one on the pelvis(hips), which had a sampling rate of 200 Hz. An example of how this looks from the right camera can be found in figure 1.

For this research only the image from the right side was used, as an image from the side contains more information about the motion, since a larger part of the dog is clearly visible. The right one was chosen as opposed to the left one, since there are people walking the dogs on the treadmill and they were mostly walking on the left of the dog, resulting in better images from the right camera.

The three different locations for the IMU sensors can have an impact on the performance of an AAR model. The IMU data from the head is, most likely, hardest to use, as a dog can and will make many movements with the head that are unrelated to the labeled activities. These movements do influence the IMU data possibly confusing the AAR model. At the whithers this effect is significantly less, but it can still have an impact. This effect was assumed to be the least at the pelvis, therefore, solely the IMU data from the pelvis was used.

The IMU sensors collected the following data:

(1) accelerometer data in x-, y-, and z-direction with a range of ± 16 g
(2) gyroscope data in x-, y-, and z-direction with a range of ± 2000 °/s
(3) high-g accelerometer data in x-, y-, and z-direction with a range of ± 100 g

To synchronize the IMU data with the video data an LED is visible on all video footage. If the LED is on, this means that the IMU sensors are recording.

3.2 Preprocessing

3.2.1 Synchronization. The first step in preprocessing was to synchronize the IMU data with the video data. For this all videos from the right camera were trimmed to only the frames where the LED is on, and thus the IMU sensors are recording. This was done by manually selecting the first frame where the LED was on and the last frame the LED was on, and then cutting away the parts of the video before and after these points.

Next, the IMU data was resampled to match the framerate of the videos, using the scipy.signal.resample() method. This resulted in an IMU sample for every frame of the video.

3.2.2 Labeling. The second step is to provide labels for the IMU data and the video data. This dataset was created with the dogs performing 2 activities: walking and trotting. All segments in the videos with these activities were labeled as such. However, these activities were not performed by the dogs constantly. There were some segments in the video with transitions between activities and other activities, like shaking with the head, that were not labeled. There were also was one other activity present that occurred frequently, namely, standing. Therefore standing was included as a label and the final dataset contains 3 activities.

3.2.3 Dog detection. Before the images were placed into the neural network, first the dog was extracted from every frame of the video. This is to make sure that mainly the relevant information is used for training the neural network. During this process, some images were discarded because the dog could not be recognized (entirely) in them.

3.3 Generating IMU data

For the first part of answering my research question, an end-to-end deep learning model was developed to predict IMU data from video. For this, the image regression of AutoKeras was used. The inputs for the training were the video frames and IMU data (as ground truth). The model was tasked to use one frame to generate one matching IMU sample. For the IMU data, it was decided to directly generate the magnitude of the 3D acceleration vector, which is calculated by:

$$magnitude = \sqrt{x^2 + y^2 + z^2}$$

This was done, as generating the x-, y-, and z-components of the acceleration separately would most likely accumulate a larger error when calculating the magnitude.

The quality of prediction is determined by a loss function, which is a number calculated from the output, of which the goal is to minimize or maximize it. In this research, the mean square error(MSE) was chosen, as the goal would be to get the predicted value as close to the real value as possible per frame. If temporal data would be used to generate IMU data, it could be beneficial to take this into account in the loss function, however, this is not the case in this research.

To keep training times reasonable the images were re-scaled to 32×32 pixels. The number of trials was limited to 10, so 10 different models were allowed to be tested. The training was done in 2 epochs, so the training set was passed through the model twice. After the training, the model was reviewed by comparing the predictions based on the given video data to the known IMU data. After that, the model was used to generate IMU data for training the animal activity recognition model.

3.4 Animal activity recognition

For the second part of the research, multiple classification models were trained. For this a naive Bayes classifier was used, as this has proven to be a reliable classifier for animal activity recognition using IMU data [7, 8]. The models were trained using different training sets, 3 with only real IMU data, 2 with a combination of real and generated IMU data, and 1 with only generated IMU data.

For the training of the AAR model, a window size of 2 seconds was used. This means that the features are determined based on 2 seconds of IMU data and that every window of 2 seconds has a single label. In addition to the 2-second window 50% overlap between windows was used. These characteristics were chosen as they are commonly used and effective.[8, 11]

The features that were used for the classification were based on accelerometer data. First, the magnitude of the acceleration was calculated from the accelerations in the x-, y- and z-direction. Then, the data was normalized. Finally, for the magnitude, the mean, median, standard deviation, minimum, maximum, 25^{th} percentile, and 75^{th} percentile were calculated. These features were chosen as they have proven to be effective in previous research.[7]

The metrics of these models were then determined by test data so that the accuracy scores and F1 scores can be evaluated.

4 RESULTS

4.1 IMU data generation

For the IMU generation data from 2 dogs (Colin and Scrufy) was used. After detecting the dog in the frames and discarding invalid frames, 12830 frames remained to train the models on. The results of training models using AutoKeras gave a resnet50-based model, with a mean square error(MSE) of 73.1 as the best model.

Resnet50 is a so-called residual neural network, which is a CNN that uses residual learning to overcome the issue of degrading/exploding gradients.[5]

One on the most important reasons that this model was selected by AutoKeras, is most likely that it is among the first models that are attempted to train. This would be the case since resnet50 and similar models are some of the state-of-the-art CNN models[1].

In figure 2 a set of generated IMU data is plotted against a set of real IMU data and in figure 3 the first 500 frames of the same fragment can be seen. The MSE is, in this case, a measure of the average distance between the generated and real IMU data. The higher the MSE value, the worse the prediction is, so 73.1 would not be a desirable result.

Though locally some small resemblances between the real and generated IMU data can be found, it seems that the model is not very successful in predicting the IMU data. Especially not compared to previous results, where the generated IMU data would clearly follow a similar trajectory as the actual IMU data [9, 11, 15].

This shows that replicating a single IMU data sample from a single image is not doable in this configuration. It could be for several reasons. First of all the dataset could be too small. The model might

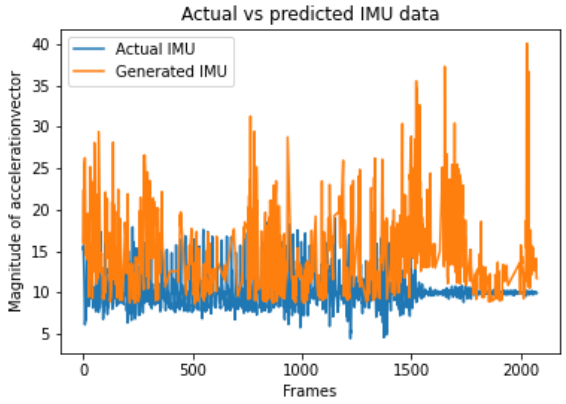

Fig. 2. Comparison of generated and real IMU data

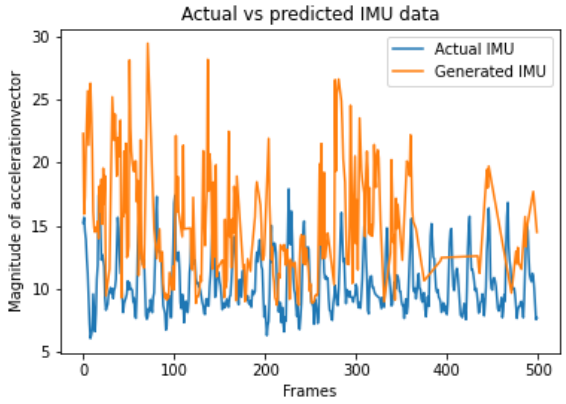

Fig. 3. 500 frames of generated and real IMU data

require more frames and epochs to achieve a lower MSE. Secondly, it could be that no truly suitable model was found yet, as only a very limited amount of trials was used. Thirdly, the low resolution could be a problem, there might just not be enough information left in the images at the 32×32 pixels resolution. Lastly, it could be that a single image is simply not enough for current neural networks to determine acceleration. While the first three reasons might have some influence, I expect that the fourth one is the main reason for the poor results in this section. This would be due to the fact that acceleration becomes visible over time and by looking at just one image at a time, the temporal information from the videos remains unused.

4.2 Animal activity recognition

4.2.1 Metrics. The most common metrics to evaluate a classifier on are precision, recall, accuracy, and the F1 score. These scores are all calculated based on the classification divided into 4 areas:

(1) True Positive (TP), is what was correctly classified as this class
(2) True Negative (TN), is what was correctly classified as not being this class
(3) False Positive (FP), is what was incorrectly classified as being this class
(4) and False Negative (FN), is what was incorrectly classified as not being this class

The accuracy is just the amount of correct classifications compared to the total amount of classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall is a metric that indicates what percentage of positive classifications were done correctly.

$$Recall = \frac{TP}{TP + FN}$$

Precision is a metric that indicates what percentage of positive classifications were actually correct.

$$Precision = \frac{TP}{TP + FP}$$

For some applications, high precision or high recall is desired. In this case, however, there is no particular need for either metric to perform well, therefore we can use the F1 score, which takes into account both precision and recall.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.2.2 Classification results. To evaluate the impact of generated IMU data, a naive Bayes classifier was trained on 6 different training sets and evaluated using the same training set. The composition of these datasets can be found in table 4.2.2. All training data came from the dogs Colin and Scrufy. The generated data was based on their images as well. As testing data, the IMU data from Sam was used.

The first thing to note about the results in table 4.2.2 is that for all compositions of the training data, the accuracy and F1 score are both pretty high. Even with only 28 training windows split over 3 classes the accuracy was able to get above 90%. Compared to other research the accuracy and F1 score were higher [9, 11]. Part of this can most likely be attributed to the limited size of the dataset used in this research, especially the fact that there were only 3 activities.

Looking at the metrics for the training sets that included generated IMU data in comparison to the ones only containing real data, it seems that generated IMU data can bring some improvement to the metrics, as all training sets containing generated IMU data outperformed the ones that did not. The most interesting thing to note might be that 28 windows of generated IMU data are more suitable to train the classifier than 28 windows of actual IMU data.

The results that were achieved with generated IMU data seem very good compared to other research, especially when looking at the performance of generated IMU only.[9]

It seems that even though the generated IMU and the real IMU plotted against one another do not show too much resemblance, the information required for the classifier is extracted from the images. The information in the generated IMU even seems better than the actual IMU it was supposed to replicate.

Since the AAR model does not directly use the IMU data, but features based on the IMU data, it could be that even though the generated IMU data does not resemble the actual IMU data, the features remain roughly the same. This could be an explanation for a similar performance of the AAR model. However, it is harder to find a reason why the generated data would have a better performance. A possible explanation could be that due to inaccuracy of the IMU generator, certain features become more clear in the data, resulting in features that help better distinguish the activities.

5 CONCLUSION

In this paper, the possibility of using neural networks for generating IMU data for use in training AAR models was shown. Even though the dataset and the number of different activities were relatively small, there seems to be evidence that neural networks be used on video to generate IMU data in such a way that information used for classification can be extracted, as the naive Bayes classifiers that were trained with generated IMU data outperformed the ones trained or real IMU data.

6 FUTURE WORK

Based on this research there are many directions to improve or extend in. First of all, it would be interesting to see, if the performance holds up when the size of the dataset increases, particularly when more different activities are added.

Secondly, it would be interesting to see what models AutoKeras can create given more training data and more time, as the use of AutoKeras in this research was quite limited.

Lastly, it might also be interesting to see what deep learning models like long short-term memory(LSTM), which can make use of temporal information, can do in generating IMU data. As much information about acceleration could come from the difference between frames as opposed to the situation in a single frame. It was discussed in section 4.1 that lack of temporal information is one of the most likely reasons for the high MSE.

REFERENCES

[1] Yanjiao Chen, Baolin Zheng, Zihan Zhang, Qian Wang, Chao Shen, B Zheng, Z Zhang, Q Wang, and Qian Zhang. 2020. Deep Learning on Mobile and Embedded Devices: State-of-the-art, Challenges, and Future Directions. *Comput. Surveys* 53, 4 (2020). https://doi.org/10.1145/3398209

[2] Rafael Cuperman, Kaspar M.B. Jansen, and Michał G. Ciszewski. 2022. An End-to-End Deep Learning Pipeline for Football Activity Recognition Based on Wearable Acceleration Sensors. *Sensors* 22, 4 (2 2022). https://doi.org/10.3390/S22041347

[3] Odongo Steven Eyobu and Dong Seog Han. 2018. Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable IMU Sensor Data Using a Deep LSTM Neural Network. *Sensors 2018, Vol. 18, Page 2892* 18, 9 (8 2018), 2892. https://doi.org/10.3390/S18092892

[4] Puja Gupta, Varsha Sharma, and Sunita Varma. 2022. A novel algorithm for mask detection and recognizing actions of human. *Expert Systems with Applications* 198 (7 2022), 116823. https://doi.org/10.1016/J.ESWA.2022.116823

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. *Deep Residual Learning for Image Recognition*. Technical Report. http://image-net.org/challenges/LSVRC/2015/

[6] Haifeng Jin, Qingquan Song, and Xia Hu. 2019. Auto-Keras: An Efficient Neural Architecture Search System. (2019). https://doi.org/10.1145/3292500

[7] Jacob W Kamminga, Duc V Le, Jan Pieter Meijers, Helena Bisby, Paul J M Havinga, N Meratnia@utwente Nl, ; Paul, J M Havinga, and Nirvana Meratnia. 2018. Robust sensor-orientation-independent feature selection for animal activity recognition on collar tags. *dl.acm.org* 2, 1 (3 2018), 27. https://doi.org/10.1145/3191747

[8] Jacob W. Kamminga, Nirvana Meratnia, and Paul J.M. Havinga. 2019. Dataset: Horse movement data and analysis of its potential for activity recognition. *DATA 2019 - Proceedings of the 2nd ACM Workshop on Data Acquisition To Analysis, Part of SenSys 2019* (11 2019), 22–25. https://doi.org/10.1145/3359427.3361908

[9] Hyeokhyen Kwon, Catherine Tong, Yan Gao, Gregory D Abowd, Nicholas D Lane, Thomas Plötz, and Harish Haresamudram. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *dl.acm.org* 4, 3 (9 2020), 87. https://doi.org/10.1145/3411841

[10] Hiroki Ohashi, Mohammad Al-Naser, Sheraz Ahmed, Takayuki Akiyama, Takuto Sato, Phong Nguyen, Katsuyuki Nakamura, and Andreas Dengel. 2017. Augmenting Wearable Sensor Data with Physical Constraint for DNN-Based Human-Action Recognition. (2017).

[11] Vitor Fortes Rey, Kamalveer Kaur Garewal, and Paul Lukowicz. 2021. Translating videos into synthetic training data for wearable sensor-based activity recognition systems using residual deep convolutional networks. *Applied Sciences (Switzerland)* 11, 7 (4 2021). https://doi.org/10.3390/app11073094

[12] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let there be IMU data: Generating training data for wearable, motion sensor based activity recognition from monocular RGB videos. *UbiComp/ISWC 2019- - Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (9 2019), 699–708. https://doi.org/10.1145/3341162.3345590

[13] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (12 2017), 843–852. https://doi.org/10.1109/ICCV.2017.97

[14] Daniella Teixeira, Simon Linke, Richard Hill, Martine Maron, and Berndt J. van Rensburg. 2022. Fledge or fail: Nest monitoring of endangered black-cockatoos using bioacoustics and open-source call recognition. *Ecological Informatics* 69 (7 2022), 101656. https://doi.org/10.1016/J.ECOINF.2022.101656

[15] W Vanwinsen. 2021. Evaluating the performance of simulated IMU data for animal activity recognition. (2021).

[16] Xin Wan. 2021. Generating IMU Data for Horse Activity Recognition from Horse Videos. (2021).

	204 real	204 real + 28 generated	28 generated	28 real	28 real and 28 generated	56 real
Accuracy	0.951	0.975	0.975	0.906	0.970	0.936
F1 score	0.950	0.975	0.975	0.901	0.970	0.935

Table 1. Overview of achieved accuracy and F1 scores

	Activity	Standing	Walking	Trotting	Total
204 Real	Real	43	105	56	204
	Generated	0	0	0	0
204 Real + 28 Generated	Real	43	105	56	204
	Generated	9	13	6	28
28 generated	Real	0	0	0	0
	Generated	9	13	6	28
28 real	Real	6	14	8	28
	Generated	0	0	0	0
28 real and 28 generated	Real	6	14	8	28
	Generated	9	13	6	28
56 real	Real	12	29	15	56
	Generated	0	0	0	0
Test	Real	20	20	41	81

Table 2. Composition of training data