UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Comparison of Spatio-Temporal Hand Pose Denoising Models

Author:
Johnny NÚÑEZ

Supervisors:
German BARQUERO, Sergio ESCALERA, Cristina PALMERO

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

July 15, 2022

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Comparison of Spatio-Temporal Hand Pose Denoising Models

by Johnny NÚÑEZ

Human pose is present in computer vision digitalization, but errors always occur in data capture. As a result, these are very frequent in datasets used for other research lines, particularly for parts of the body that are moving continuously, such as the hands. Such undesired noise can be extremely harmful to many downstream tasks like human behavior analysis, forecasting, or action spotting, among others. In this thesis, we present a systematic comparison of several state-of-the-art approaches for hand pose denoising on video sequences. Such methods are also evaluated in the frequencial space, which have provided important performance boosts in other closely related fields. We use the hand annotations available in a recently released dataset: UDIVA v0.5. Additionally, we present an innovative perturbation algorithm to generate noisy hands while maintaining hand structure and coherence. Our results demonstrate that our approach succeeds in denoising the hands' trajectories, unlike other state-of-the-art studies which do not address this issue. We also achieve very satisfactory results in human pose denoising. This opens new challenges for model improvement on extremely noisy and occlusion pose cases.

Acknowledgements

I would like to express my deep and sincere gratitude to my research supervisor, Dr. Sergio Escalera, MSc. Germán Barquero and the future Dra. Cristina Palmero. Dr. Sergio Escalera, Professor and Head of the Human Pose Recovery and Behavioral Analysis Research Group at the University of Barcelona, Computer Vision Center. His dynamism, vision, sincerity and motivation have inspired me deeply. They have taught me the methodology to carry out research and to present research papers as clearly and timely as possible. It has been a great privilege and honor to work and study under their guidance. I am extremely grateful for what you have offered me, for giving me the golden opportunity to do this wonderful project on the topic "Hand Pose Denoising", and especially to Cristina for her magnificent dataset and be my role model. I would also like to thank them for their friendship, empathy, and great sense of humor during this thesis and for being able to attend the HupBa Group seminars.

I would like to dedicate some sincere words to Germán Barquero, a diamond in the rough, for all the help and doubts resolved throughout this project. I wish you the best of luck in your PhD, you will be a great doctor someday.

On the other hand, I wanted to thank my family, who has always supported me and helped me to be the person I am now. I wanted to make a special dedication to my darling, Paula Sanchez, Primary School Teacher, for her unconditional support, and for putting up with me in the difficult moments during the thesis.

Contents

Chapter 1

Introduction

In computer vision, human pose estimation consists in determining the position of a person or an object from an image or a video. These poses are represented through keypoints (also referred to as landmarks or joints) in image space, and connected together, resembling the structure and symmetry of the human body. This is useful to simplify how a computer understands how to identify a person's posture, both the position of the human pose and the orientation relative to the camera in relation to a given person or object. This technique is widely used in applications such as the metaverse, video games or movies, where it greatly facilitates the work of the creators of this type of content. The current methods are based on identifying and estimating the human pose given an image, but the most recent methods have been extended to video, where it is a much bigger problem since you get a historical context of the human pose through a sequence of images. These current methods are not perfect, which leads to noise generation, for example, a reference point being far away from the reference point in image space. There are also problems in reference to the reconstruction of the human pose, obtaining incomplete skeletons, according to this problem, we need methods that fix these poorly estimated or noisy poses. So, pose estimation is linked to a process of pose refinement or pose denoising.

There exists a wide range of datasets that have human pose information (see Barquero et al., 2022b for a review), not only through audiovisual data but also through labeled data by humans or data extracted from automatic algorithms, such as landmarks, gaze direction, 3D shapes, etc. Some of the datasets that contains human pose are: Johnson and Everingham (2010), Ionescu et al. (2014), Lin et al. (2014), Li et al. (2019b), Marcard et al. (2018), Andriluka et al. (2014), Güler, Neverova, and Kokkinos (2018), and Zhang et al. (2019). These datasets containing automatic extractions may themselves have limitations by the estimation algorithm itself because pose estimation may not be consistent. Therefore, the refinement of poorly detected or poorly constructed poses is needed.

Unfortunately, the way to solve this challenge is to hire human annotators and fix the incorrect poses manually through visual inspection, which is time-consuming and potentially inconsistent, since each person may have different opinions about a poorly performed pose. Another solution could be the use of interpolation techniques, to approximate poses through other poses, but such interpolations may have a very limited application, since they only estimate an unknown value between two or more known keypoints. This allows us to model or apply smoothness to abrupt changes, therefore it is only allowed to be used with already known data, and not with occluded human poses for example. When estimating a pose, estimation errors often occur because the person does not appear in the image, or appears partially. In videos, it often happens that there may be segments where the part of the body to be considered is totally occluded, losing the information of the temporality in that segment, giving way to the generation of totally stochastic poses. There are cases where

more than one object or person appears in the image, this can confuse the estimation algorithms in assigning reference points of the object or person, deforming the human pose. Therefore, sufficiently robust and fast algorithms are needed to improve all existing solutions.

Recent studies (Mao, Liu, and Salzmann, 2020) indicate that the use of Deep Learning is a solution to consider, both for static images and videos, being able to find solutions through the spatial information of the image itself or with temporal information, by learning from the past and the future of a corrupted segment. The objective of this thesis is to find a solution for 3D hand landmarks, because it is the noisiest part of the human body, as it is constantly moving and difficult to detect using pose estimation algorithms. The study is performed on automatically extracted landmarks. Our work is evaluated on the UDIVA v0.5 dataset (Palmero et al., 2022), a subset of the UDIVA dataset (Palmero et al., 2021). This dataset contains audiovisual clips, transcripts, and metadata, as well as automated annotations of 145 face-to-face dyadic interaction sessions. 134 participants participated in these sessions, which are divided into conversational, collaborative, and competitive components, with diverse hand movements and hand-object/human interactions. Existing works (Barquero et al., 2022a) have explored the effect of noise on this dataset in other applications such as forecasting, and it has proven to be a detrimental limitation for predicting the future, since adding only 10% of noisy hands annotations resulted in a drop in performance in most of its models. Despite the fact that the dataset was subject to manual corrections, only 7% of the errors were fixed. Finding automatic methods that can correct all miscalculated poses is still quite difficult.

Therefore, we propose fixing corrupt hands pose segments on UDIVA v0.5 through recent deep learning algorithms that help to improve the dataset and its usefulness. In particular, we evaluate Multi-Layer Perceptron (MLP) models in original space and frequential space (using the Discrete Cosine Transform), sequence-to-sequence (Seq2Seq) recurrent models (Sutskever, Vinyals, and Le, 2014), Transformers (Vaswani et al., 2017). As one of our main contributions, we propose mixed versions of the aforementioned models. All approaches are evaluated assuming that we are provided with a temporal window of observed behavior (observation window) in order to denoise the middle frame of the observation window. Furthermore, we develop a perturbation generation algorithm that coherently maintains the hand structure and motion over a sequence of frames. Finally, we address the problem of the correctness of the pose (how the hand is placed), and the problem of the trajectory (where our hand is placed).

The remainder of this thesis is organized as follows. First, we describe the latest works about pose refinement or pose denoising (chapter 2). Then, we describe the data used (chapter 3) and an intelligent perturbation hand algorithm. Next, we explain the recent architectures in deep learning to apply to this challenging research (chapter 4). Furthermore, we describe the experimental evaluation and discuss the results obtained (chapter 5). Finally, chapter 6 concludes the thesis.

Chapter 2

Related Work

Pose denoising is still a topic of discussion nowadays since pose estimation algorithms are not perfect. The input of most of these algorithms usually consists of static images. In these cases, the pose is retrieved without exploiting the temporal dimension. Many other applications are based on these automatically extracted low-level representations, such as body or face landmarks. In behavior forecasting, for example, previous body landmarks are used to predict where the landmarks will be located in the future. Since these methods often struggle when dealing with noisy pose sequences (Barquero et al., 2022a), some methods have proposed pose refinement strategies (Moon, Chang, and Lee, 2019). When dealing with temporal information, such as videos, a good approach is to use recurrent neural networks (Kim and Chang, 2021) in order to obtain the information along the sequence, especially since the problem is more complex than it appears, as maintaining the pose trajectory is required, in addition to modeling the pose position itself.

Despite the different perspectives mentioned above, there are some mistakes associated with human poses. The most common errors are displacement of a keypoint relative to the ground truth, known as noise, occlusion of keypoints, mistaken inversion of keypoints with respect to the same person, and confusion of keypoints between different people co-appearing in the scene, or in the same person, when the left hand is confused with the right hand and vice versa. As such, recent studies using low-level representations to solve these common problems are of interest. Occlusion is one of the most serious cases that can arise. Carissimi et al. (2018) presented a method to predict missing 2D joints from incomplete human body poses. A pose estimation method based on a simple AutoEncoder (Rumelhart, Hinton, and Williams, 1986) is used to resolve the problem of missing joints.

Working with videos increases the complexity of the problem to be solved. Fortunately, the exploitation of the temporal information allows one to leverage past and future to obtain a context about the pose to be corrected. Wang et al. (2021) proposes a new deep network STRRN (Spatial, Temporal, and Residual Network) with a batch prediction method, which is capable of predicting a large number of frames at once, allowing one to make use of long-term time-based objective functions in order to accurately learn motion multimodality and variance to resolve a motion prediction problem. It makes use of Bidirectional Recurrent Networks (Schuster and Paliwal, 1997), in specific, Long Short-Term memory (LSTM) (Hochreiter and Schmidhuber, 1997), to learn on the time axis, using the sequence-by-sequence approach, and additionally an encoder/decoder to force reconstruction of information at each step. In the iterative motion prediction process, they observe periodic jumps. They propose to approach the removal of this high-frequency noise as a learning problem in order to generate a signal to cancel the noise, so they add a residual layer. A similar idea is proposed by Cui et al. (2019) where bidirectional recurrent layers are used as well. Their neural network is called the Bidirectional Attention Network (BAN), in which

they embed an attention layer into the recurrent layers which can capture long-term dependencies and adaptively extract relevant information at any given time step. The network explicitly selects the relevant context of a damaged motion frame in order to repair it. Specifically, the elimination of noise from hand motion (Zhou et al., 2021), and shows an algorithm on how to add perturbations in the hand, and a Graph Auto-Encoder neural network (Scarselli et al., 2009), to preserve the hand structure in each frame throughout the sequence.

On a different note, there also is the non-deterministic conception of the denoising problem. It is particularly important for large segments with wrong or missing poses, where many possible distinct sequences of poses are equally valid and plausible. Gu, Zhao, and Zhang (2022) propose a Conditional Variational Autoencoder (CVAE) within the framework of a recurrent neural network, in which Gaussian noise is introduced to ensure probabilistic generation. Additionally, they introduce a regularization loss in order to explicitly promote the diversity of the generated samples.

The latest studies show a new way to eliminate noise in sequences using the frequency space by means of the Fourier transform. This allows to solve the problem through the frequency domain and identify the high and low frequencies to eliminate noise. For example, jitter can be understood as a high-frequency signal over temporal information. In the works of Lin and Lee (2019), Mao et al. (2019), Mao, Liu, and Salzmann (2020), Katircioglu et al. (2021), and Gauss et al. (2021), the DCT was used, reducing the original space to only a coefficient matrix in order to resolve the issue of trajectory. Therefore, most of the error originates from the first DCT coefficients, which have low frequencies.

The conclusion is that while few studies have suggested solutions to correct hand poses, we believe that they can reduce the noise in hand poses. All studies have been related to body landmarks except the one from Zhou et al. (2021). With respect to architectures, the good performance of RNNs suggests that leveraging temporal information is useful as long as we have a coherent sequence of data, since a minimum perturbation can destroy the refinement of our sequence pose. Attention mechanisms are a good solution to obtain the context and retrieve the most relevant information to reconstruct the pose, and residual layers can help to smooth the final pose prediction. Furthermore, recent works have highlighted the benefits of considering the pose in the frequential space. Such strategies help the refinement of both low and high pose frequencies. We also feel it could be a great option and offer a good solution to our problem.

Chapter 3

Data

3.1 Dataset

In this work, we use UDIVA v0.5 (Palmero et al., 2022), a dataset composed of 145 dyadic interaction sessions divided into 4 different tasks each: Talk, Lego, Ghost, and Animals. This dataset provides automatically extracted body, face, and hand landmarks, as well as 3D gaze direction vectors. Specifically on hands, FrankMocap (Rong, Shiratori, and Joo, 2021) was used to extract 3D hand landmarks, because it infers fairly accurate hand landmarks in the recurrent scenario where hands are interlaced, interacting with objects or mildly occluded. SiamRPN++ (Li et al., 2019a) was used to track hand detections required for landmark extractions when spatial or temporal gaps were observed. This data is available for two views, each focused on a single participant. In this section, we will describe the dataset used and the processing applied to the data, including the data perturbation algorithm used and the application of Gaussian noise on the hands.

3.1.1 Data

First of all, we need data to train, data to validate our training and finally data to test our trained models. In this case, UDIVA v0.5 contains 116 sessions and 99 participants for training, 18 sessions and 20 participants for validation, and 11 sessions and 15 participants for testing.

Specifically, Palmero et al. (2022) indicates that the automatically extracted pose annotations from the validation and test sessions were revised by humans to assess their correctness and improve, if possible, some of the wrongly extracted human poses. The hands annotations may exhibit a number of related problems, such as left-right hand mismatch or false positives or negatives. Additional labels were provided by the manual revision process to distinguish between these instances, as shown in Figure 3.1, examples of each label annotated by the reviewers. First, if the annotations' orientation was right and the location of the fingers matched their anatomical positions, the quality flag was set to *correct*. If the position or shape of the fingers did not match, but the general orientation of the hand was correct, they were classified as *mild*. Otherwise, the hand quality was classified as *severe* if neither the hand orientation nor the fingers could be correctly inferred. A visibility flag was also used for each hand, set to *visible* when the hand was visible and *not visible* when it was occluded. The latter usually happens when the person placed their hands under the table or one of the hands was occluded by an object. In addition, cases where the left hand was detected as the right hand or vice versa (hand *switch*) were annotated even when only one hand was detected. To increase the proportion of accurate hand annotations, the annotation reviewers searched for sequences of frames that had no movement or extremely slow hand movements and contained

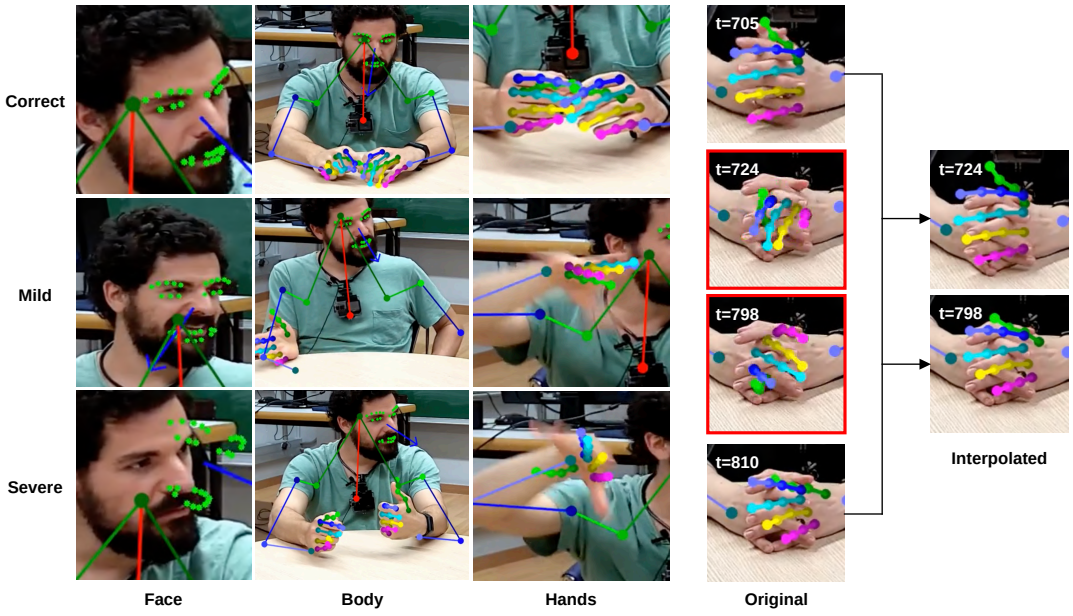

FIGURE 3.1: In (a), examples of *correct*, *mild* and *severe* quality labels for face, body and hands landmarks. In (b), a sequence of 104 frames with wrong left-hand landmarks which was fixed by linearly interpolating the landmarks from the last and the first correct extractions before (frame 705) and after (frame 810) the sequence, respectively. Image from Palmero et al., 2022

wrong pose estimations. They attempted to maximize the number of correct hand annotations by identifying sequences of consecutive frames $(t_0, ..., t_n)$ with $n \geq 2$, no or very slow hand movements, and correct hand landmarks at t_0 and t_n, but with wrong or missing hands in all frames in between $\{t_i\}_{1 \leq i \leq n-1}$. On the basis of those segments, segments for which a linear interpolation between the landmarks (t_0, t_n) could produce valid hands for the interval (t_0, t_n) were identified (more information in Palmero et al., 2022). From these, those sequences that could be fixed by simply linearly interpolating the hand pose between two selected frames were selected as fixed by interpolation (FBI). The interpolation was then performed. Such scenarios are mostly segments with mild occlusions or hand-to-hand interactions where the hand remained static for a while. Unfortunately, the hand annotations are highly noisy (around 20% of the frames had wrong hand annotations). Even after the manual hand interpolations, the refined annotations for the validation and test sets only include manual corrections of up to 7%. These hands correspond to frames where there are occluded hands, interlaced hands, handshakes, off-camera hands, object-hand interactions, or extreme hand poses.

Table 3.1 summarizes the results of the manual revision for the test and validation sets in UDIVA v0.5 talk sessions. The hands linear interpolation was extremely beneficial, since it was capable of repairing up to 7% of the annotations in both validation and test sets (1714 full segments for both hands with an average length of 34.5–78.3 frames). Between 80% and 90% of accurate hand annotations are found in the validation and test sets. In conclusion, the interpolated frames allow us to create a ground truth that can be used to teach our algorithms how to correct wrong hand pose extractions. We have the same original and corrected pose. However, such set of pairs of wrong and ground truth poses may not suffice to build our denoising

model. Fortunately, techniques like data augmentation can compensate for the lack of data. More details are given in the following sections.

	Validation/Test sets					
	Correct	**Mild**	**Severe**	**Switched**	**VF**	**FBI**
Left hand	88.9/87.0*	7.2/8.0	3.9/5.0	0.4/0.2	2.6/1.6	6.6/7.1
Right hand	90.1/82.3*	5.4/9.7	4.5/8.0	0.4/0.2	2.0/3.4	6.9/7.0

TABLE 3.1: Visual inspection process for validation and test sets: prevalence of each label (% of frames). Abbreviations: VF, Visibility Fixed; FBI, Fixed By Interpolation. * Including switched, VF, and FBI annotations. This table is taken from Palmero et al. (2022).

3.1.2 Data splits

To train and evaluate our denoising models, we generated new training, validation, and test splits from the original UDIVA v0.5 validation and test sets. This was done by: first, ensuring that there were no shared individuals between splits; second, trying to maximize the number of unique subjects in the validation and testing sets; and finally, making sure to keep a high number of interpolated segments in the test set. As a result, the following frame percentages are obtained:

- Training: 65% including 19 sessions, and 19 different subjects.
- Validation: 17% including 5 sessions and 7 different subjects.
- Test: 17% including 5 sessions and 9 different subjects.

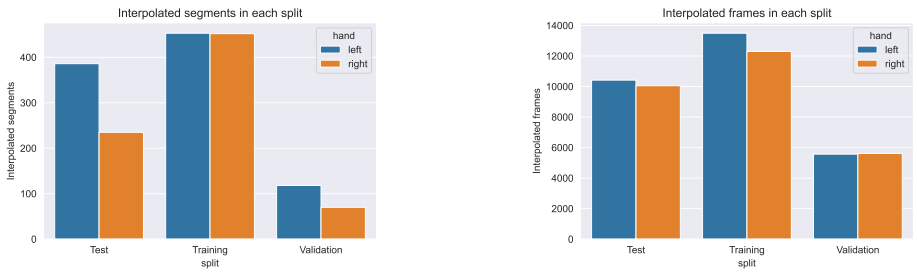

FIGURE 3.2: New training, validation, and test distribution of interpolated frames.

In figure Figure 3.2, we talk about frames as the sum of interpolated frames in ALL segments, while the other figure is segments where interpolation has been applied. The revised data distribution in Figure 3.2 leads to a slight imbalance as for the hands in the interpolated frames, which is more evident in the test. This does not really matter to us, because we finally performed a left hand flip.

3.2 Perturbation

As we have discussed above, the data available in the original dataset may be too scarce to train models on. In this work, we propose a method to augment the amount

of wrong hand pose extractions with available ground truth while still keeping the noise distribution present in the original dataset.

The goal of creating a disturbance is to be able to make the hand worse in order to reach more extreme cases. To start, perturbing a hand in the temporal space is not an easy task. As requirement, the physics of the hand must be preserved and it is essential that the hand remain coherent throughout the sequence.

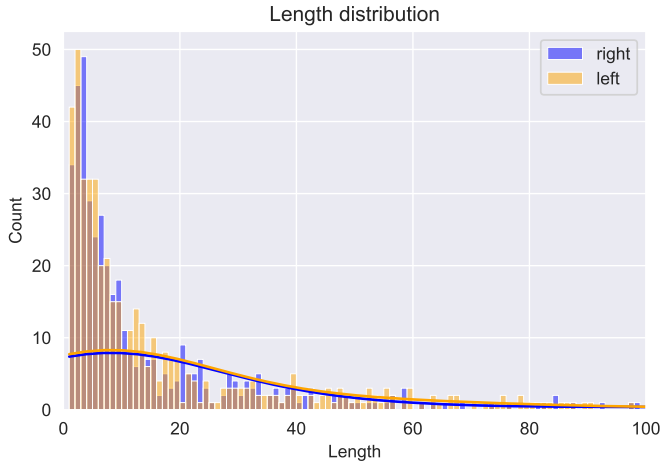

FIGURE 3.3: Histogram of corrupted segment length.

First, we analyze the amount of corrupted segments that we have in our data as a function of the length of such segments. In Figure 3.3, the highest concentration of corrupted segment frames is usually less than 20 consecutive frames, but we can still observe segments with a longer length of 100-200 frames. Finally, these results are fitted with a log-norm distribution to be able to sample the length of a noise segment.

Next, we must analyze the prevalence of segments with wrong hand pose estimations in order to replicate those in our perturbed dataset. It is possible to identify these erroneous segments because the dataset has this information (FBI labels). Our analysis showed a probability of left 0.34% and right 0.31% of each frame being the starting point of a segment of wrong pose extractions.

Then, we analyze how much noise should be applied to ensure that the hand becomes noisier but still a coherent and plausible wrong pose estimation. Given the algorithms used for hand pose estimation (3D model fitting), we can assume that there is a similarity transformation (rotation + translation + scaling) that transforms the noisy annotations to the correct annotations. This is a strong assumption, and we acknowledge it is not completely accurate. However, it helps us to simplify the problem to a great extent. To model the noise distribution, the least-squares method is applied to the fitted raw data and the cleaned data in order to determine the distribution of the transformation parameters, resulting in an initial MSE of 2858.9 +/- 4820.2 in the original data with respect to the clean data, which implies a high error, and a final MSE of 144.3 +/- 168.0 after applying the similarity transformation (3 values for rotation, 3 for translation, and 3 for scaling). However, we must point out that there may be finger bending, for example, that we are not modeling to keep the method simple. In order to account for such noise, up to some extent, we perturbate the fingers landmarks with Gaussian noise with mean 0 and standard deviation of 1 once the hand has been normalized, this allows changing the values of the coordinates of the hand to use a common scale, without distorting the differences in the

value ranges or losing information. Finally, we apply gaussian noise by applying 25% of the noise to the nearest part of the hand, increasing the percentage to 100% of the noise at the tip of the finger.

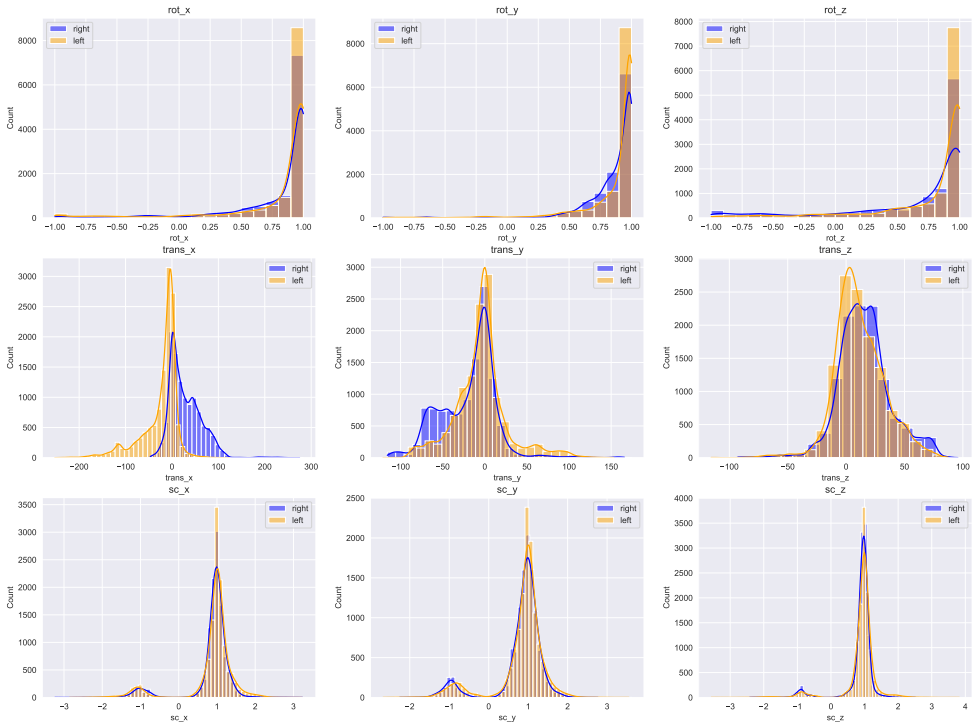

FIGURE 3.4: Distribution of noise on the 3D hand coordinates for rotation, translations, scaling extracted from the interpolated frames.

Lastly, we fit the distributions of retrieved rotation, translation, and scaling values with kernel density estimation methods, see Figure 3.4. Finally, for each hand within the perturbed segment, we sample each of the rotation, translation, and scaling values, and apply the resulting transformation. The result is our augmented noisy hand pose.

At this point, the only thing left to replicate from the original noise distribution is how many hands are missing within the segments (such landmarks are represented with 0's in the original dataset). Our analysis showed that there are 14.4% and 8.04% of missing left and right hands, respectively.

The complete algorithm is summarized in Figure 3.5.

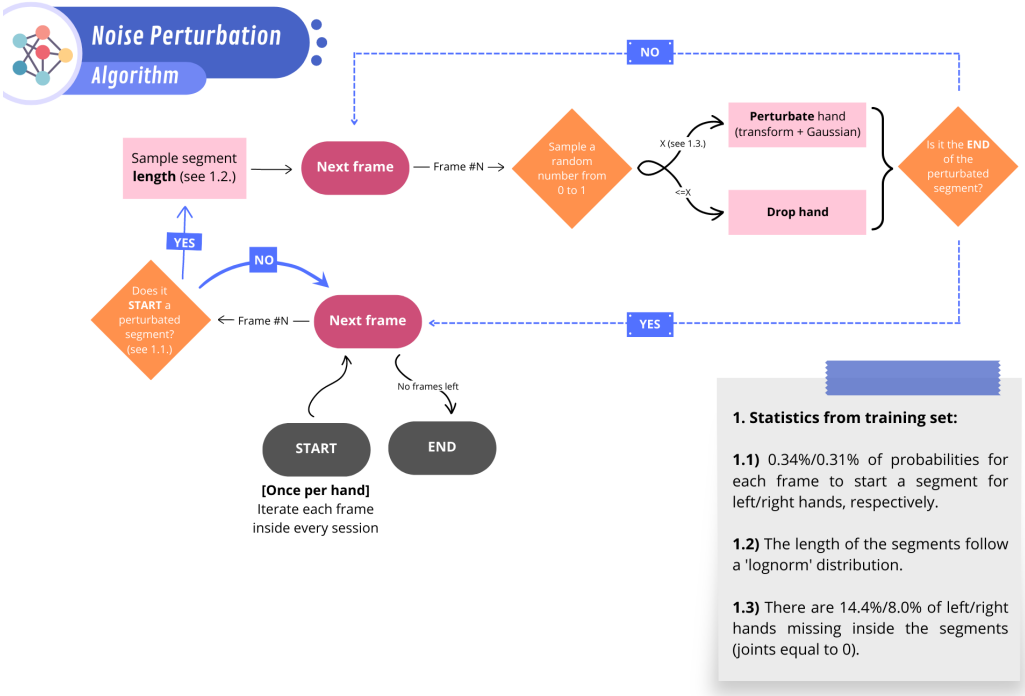

FIGURE 3.5: Summary of the noise perturbation algorithm.

Chapter 4

Methodologies

In this section, we will explain the methodology of this thesis. First of all, we will select and explain the models used and obtained from the state of the art and finally we will try to explain the proposed method.

4.1 Frameworks Overview

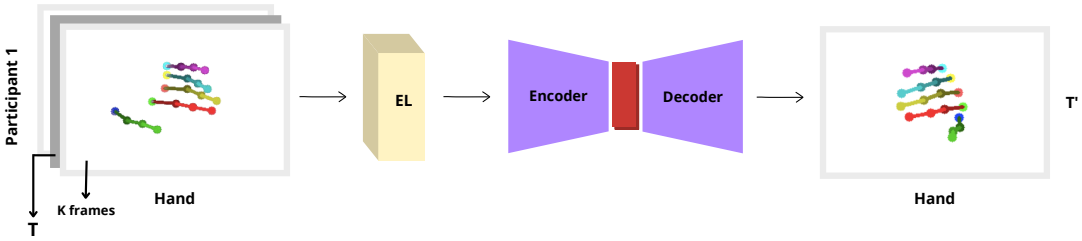

FIGURE 4.1: Overview of the general architecture used Denoising Models. Abbreviations: K, Number of observation window frames; T, Frame in the middle of the observation window; EL, Embedding Layer; T', Frame corrected

In this thesis, we propose to compare different state-of-the-art models (see Section 2) for pose refinement. First, our models need to learn from historical data, both from the past and the future. Second, an appropriate model should be able to correct noisy hands, while identifying the correct ones and leaving them untouched.

First, our frequential models are designed to incorporate a Discrete Cosine Transformation (DCT) (Ahmed, Natarajan, and Rao, 1974). As a result, our models are able to work in the frequential space, which replaces the temporal dimension. This allows us to separate the time series of each joint into frequency-spectral dimension in order to analyze the dynamics of the sequence. As a consequence, the models could learn to identify and generate low-pass filters, which consist of the elimination or attenuation of the high frequencies, when very fast movements are produced on the hand, considering this movement as noise. It is important to emphasize, Deep learning models can have problems with frequency space if the data is not formatted correctly, for instance, if the transform is not performed correctly. It is important to note that the decoder output must be able to reconstruct the data with the same dimensionality as the input data.

All models are built as encoder-decoder (Ballard, 1987) architectures (Figure 4.1) in which the hand pose is first fed into an embedding layer. This architecture compresses the input data using the encoder, resulting in a bottleneck in the middle of the model. Consequently, the model is able to identify the most important information for reconstructing and refining the pose. As for the number of K frames, we use

different window sizes 50, 100, 150, 200 to obtain more or less temporal information. Then, given an observation window, we correct the pose of the frame T in the middle of the observation window, to leverage the same length of past and future information. Finally, we have the T' pose which is the corrected T pose. The models implemented are described below:

1. A multi-layer perceptron (MLP) (Haykin, 2004) is a class of feedforward artificial neural network. MLPs consist of three layers of nodes: an input layer, a hidden layer, and an output layer. The nodes in the hidden layer process the data and transfer it to the output layer, where final results are produced. In our case, both the encoder and decoder, we used a three-layer model with non-linearities ReLU and variable dropouts depends if the frequential space or original space.

2. Seq2Seq. The idea was adopted from the original paper (Sutskever, Vinyals, and Le, 2014) and used in the papers mentioned above chapter 2 when there is temporal space. The Seq2Seq architecture consists of two parts: an encoder and a decoder. The encoder takes a sequence of data as input and encodes it into a fixed-length vector. The decoder takes the encoded vector as input and decodes it into the original sequence. Seq2Seq models are typically implemented using recurrent neural networks (RNN). In our case, we use a single Long-Short-Term Memory (LSTM) cell (Hochreiter and Schmidhuber, 1997) for recurrent models in the encoder-decoder. We also use bidirectional LSTM (Schuster and Paliwal, 1997). Finally, we use linear layers to predict the coordinates with non-linearities ReLU and dropouts of 0.5.

3. Transformer. The current state-of-the-art model (Vaswani et al., 2017) comes from Natural Language Processing (NLP). The Transformer architecture works by first encoding the input sequence into a series of vectors. The vectors are then fed into the self-attention mechanism, which generates a series of new vectors. The new vectors are then fed into the decoder, which generates the output sequence. For this thesis, we used the Spatio-Temporal Transformer proposed by (Zheng et al., 2021) whose original idea was to develop a spatial-temporal transform structure to comprehensively model human-separate relationships in each frame and temporal correlations between frames. Using the spatial-temporal transformer, future events can be predicted based on past events. Their original goal was to produce a clean 3D human pose of the center frame. In addition, we rely on the potential of this architecture tested by (Barquero et al., 2022a) for behavior forecasting. We used a single Gated Recurrent Unit (GRU) Chung et al., 2014 as the transformer model decoder.

4.2 Pose and Trajectory disentanglement

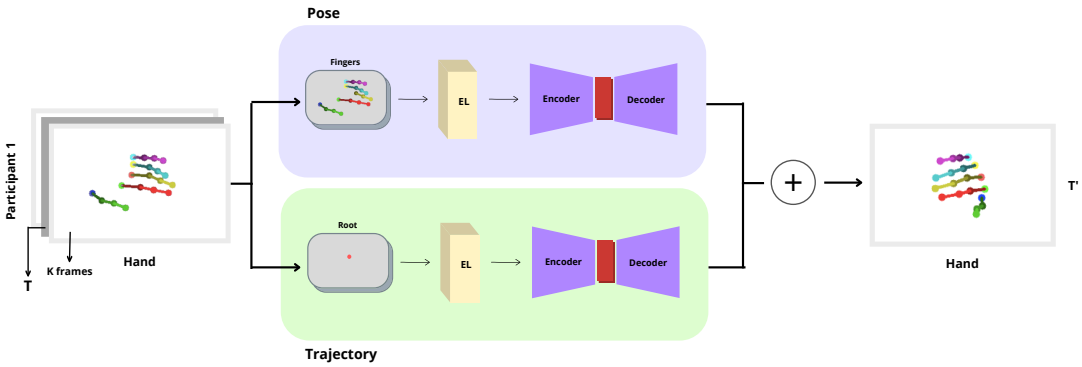

FIGURE 4.2: Overview of the architecture proposed in this thesis in all Mixed Denoising Models. Abbreviations: K, Number of observation window frames; T, Frame in the middle of the observation window; EL, Embedding Layer; T', Frame corrected

Our proposal Figure 4.2 is to solve the problem as trajectory and pose regression sub-problems. We do so by separating the reference joint considered as the root from the other joints on the hand in order to solve both problems independently. We hypothesize that a model should be able to accurately predict where the trajectory of the hand using solely information about its trajectory. Similarly, a trajectory-blinded model should be capable of fixing a hand pose from only observed past and future hand poses. For this purpose, we propose the following models:

1. Mixed-Seq2Seq. We use MLP to refine the pose of the hand which allows us to use both frequency and non-frequency space to correct the pose. While on the trajectory side, we use Seq2Seq, with LSTM and dropout 0.5. The Pose models are composed of 3 MLPs both in the encoder and decoder, with non-linearities ReLU activation and different dropouts according to the dimensionality of the model.

2. Mixed-Transformer. Similarly to the Mixed-Seq2seq implementation, Seq2Seq are used for the trajectory problem, with LSTM and dropout 0.5. To refine the pose we use the Transformer on the encoder with dropouts of 0.5. In the decoder we use Seq2Seq with GRU and 3 linear layers with non-linearities ReLU activation with dropouts of 0.5.

Chapter 5

Experiments

5.1 Experiments

In this section, we report quantitative and qualitative results of our methods, evaluated on the UDIVA v0.5 dataset. We use Hydra (Yadan, 2019) for the management and configuration of the experiments, which allows us to save and track all the data generated during training. These experiments were executed on Nvidia RTX 3090 with Pytorch 1.11 (Paszke et al., 2019), Lightning-AI 1.6.4 (Falcon, 2022) and CUDA 11.3 (NVIDIA, 2021).

5.1.1 Training data details

In this thesis, we use the 3D coordinates of the 20 hand joints provided by the UDIVA v0.5 dataset. Therefore, the information of each frame can be represented as $X_t \in \mathbb{R}^{20\times3}$. The total data consists of 19 sessions for training and 5 sessions for validation. Observations from the past and the immediate future are used to refine (or predict, in case of occlusion) the middle frame of the observation window. The observation lengths evaluated are 50, 100, 150, and 200 frames. Our ground truth is taken from cleaned data, and our observation segment is from the original (automatically extracted) data. In order to keep the data augmentation constant among all experiments, the perturbations were pre-generated for 600 epochs and applied in an online fashion during training. The algorithm used is the one shown in Figure 3.5, applied to each hand (left and right). Finally, we chose only one person to perform the experiments, out of the two available per frame.

In our work, our input consists of a sequence of frames (poses), where each frame has a 3D position of every joint. We plan to solve the problem of pose denoising from the corrupted sequence observation with bad positional landmarks or occluded hands. Due to the difficulty of the problem, we evaluate on two hands by changing the perspective of the left hand as if it were the right hand with a horizontal flip, in order to be able to use the same model for both hands. We only use the x,y coordinates of the hands since the z coordinates (depth) have not been manually verified and may be wrong and temporarily inconsistent. It is important to note that our predictions needed to be detached from the image space to compress the subspace of poses that the network needed to model. To do so, we select the middle knuckle as the center of coordinates (*root*), and subtract it from all the other joints coordinates. Therefore, our input skeleton vectors consisted of the root and 19 root-relative hand landmarks. Coordinates of missing landmarks were replaced by zeros.

5.1.2 Optimization

The training loss used was the Mean-Squared Error (MSE) in the original space, not in root-relative space, only using 2D dimensions for evaluation. All models were trained with Adam (Kingma and Ba, 2015), with a weight decay of 0.005 and a learning rate of 0.0001. The batch size was set to 64 for all models. Dropout values were tuned for each model type, with higher values for models learning from temporal space, while in models that learn from frequential space. The maximum number of epochs are 600 with an early stopping was applied to all models with patience of 25 epochs.

5.1.3 Evaluation Protocol

The evaluation was performed on the cleaned annotations of the modified UDIVA v0.5 test set (5 sessions), in which we have two options: evaluate on all frames or evaluate on those frames where there is an error difference between the raw and the cleaned ones. The left hands were flipped in axis y, before being introduced to the model for inference. Our models are evaluated with the Mean-Squared Error (MSE) and the Euclidean distance (ED) after undoing the root-relative coordinates transformation, i.e., they are computed with image coordinates. We also use the same observation window as the model itself. To evaluate all frames, we used a window stride of 1 frame, always predicting the frame in the middle of the observation. Finally, we calculate the average between the two hands to obtain the overall error value for each frame.

5.1.4 Results

We present the main results obtained in this section. First of all, we analyze the overall results and then compare the performance of models based on the different observation lengths. Then, we select the best observation length from the comparison to perform the final analysis. In summary, we have a total of 11 different models with 4 different observation window sizes, resulting in 44 experiments. In preliminary experiments we evaluated observation window lengths of 50, 100, 150 and 200 and found that 100 was consistently better, so we focus all our analysis of results on 100.

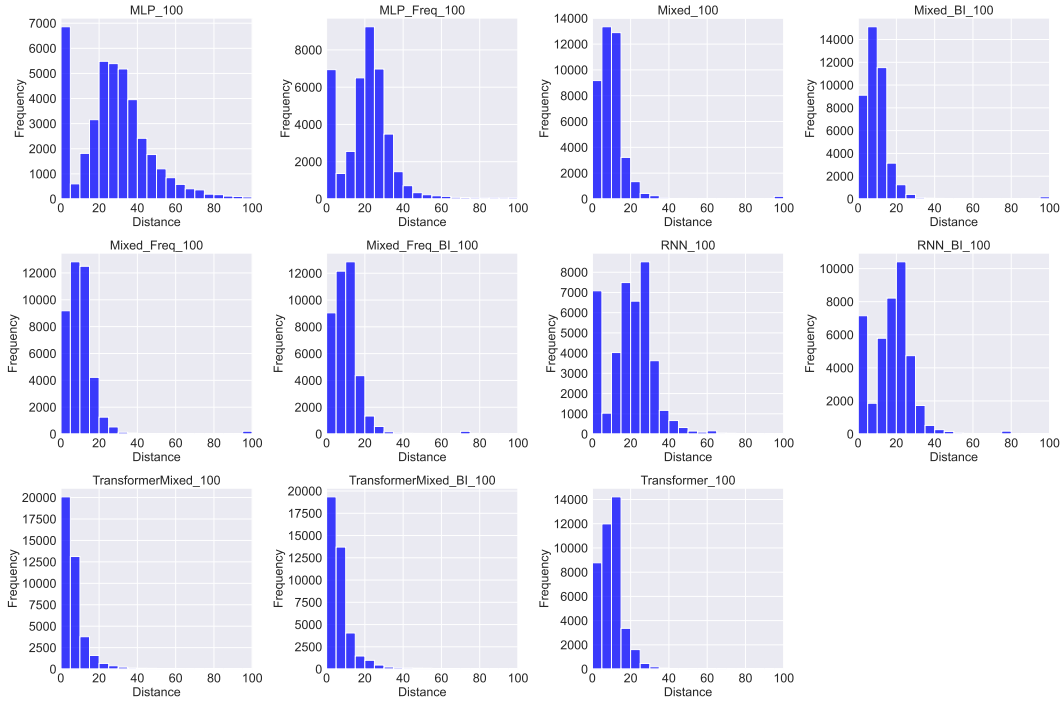

FIGURE 5.1: Error distribution in pixels for all evaluated models.

In Figure 5.1, the observations are sorted by models with 100 frames observation length. It can be observed that the models that perform the worst are those whose distribution is far from zero. This indicates that there are many errors resulting from poorly corrected frames, whereas good models are those based on a distribution close to 0. It should be noted that a column appears at a value of 0. It is evident that there are a high number of frames in which the image is left exactly as it was, that is, frames in which there were no errors between the original and the cleaned pose or images that were corrected with an exact position, getting the best possible result.

We also observe a sharper curve and closer to zero once we use models that learn on the temporal axis, such as the RNN models. In the case of the bidirectional recurrent model (BI), we achieve a little improvement over non-bidirectional ones. Finally, we have the powerful transformer architecture that has a magnificent result, Transformer and Mixed Transformer models (Freq and Non-Freq) have similar error distribution but the difference is seen with Transformer-Mixed.

Model	Original	Prediction
TransformerMixed_BI_100	9.138526	**8.649507**
TransformerMixed_100	9.138526	10.361105
Mixed_BI_100	9.138526	10.661939
Mixed_100	9.138526	11.080834
Mixed_Freq_100	9.138526	11.377095
Mixed_Freq_BI_100	9.138526	11.522652
Transformer_100	9.138526	12.230602
RNN_BI_100	9.138526	18.057422
RNN_100	9.138526	21.049572
MLP_Freq_100	9.138526	23.856506
MLP_100	9.138526	29.756484

TABLE 5.1: Mean error of the Euclidean distance in pixels of all frames under test for both the original data and the predictions, ordered in ascending prediction order.

Initially in Table 5.1, we can see that the frequential one has improved a little with respect to the non-frequential one, but not by a large margin. Therefore, we propose to treat the problem in 2 tasks, the trajectory problem and the pose problem. This is because the error may be preceded by the sensitivity of predicting the misplaced hand, i.e., if we place the hand at a considerable distance from our ground truth the error grows exponentially. So, the following results improved considerably, obtaining better results than the normal recurrent models and the temporal transformers, treating the trajectory with recurrent architectures. Finally, we obtain the best architecture using Transformers for the pose, while keeping an independent encoder and decoder for the trajectory.

Denoising Capabilities

The following analysis is provided in Figure 5.2. For ease of analysis, we choose models with an observation length of 100. On the Y-axis, from the MSE calculated by frame, we quantify how many frames have been equal to or less than the X MSE threshold, this gives us a value to measure the performance for the comparison of each model. The maximum threshold used is 10,000 MSE, while the minimum is 0. Then, the values were normalized to the maximum value to obtain a more coherent curve and the area under the curve (AUC).

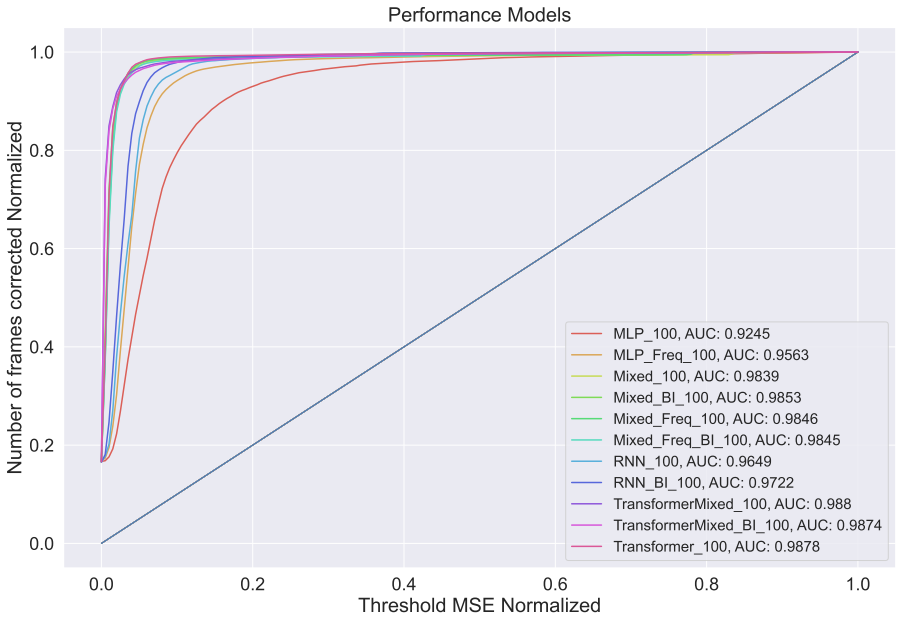

FIGURE 5.2: Performance of the 100 observation length models. The metric used is MSE, normalized between 0 and Maximum value prediction to obtain better visualization

Now we can confirm the results mentioned in the previous section. A slight improvement in frequency (AUC:0.9563) vs. non-frequency (AUC:0.9245) vs is obtained. Using models that learn from original temporal space helps to improve performance, while using mixed models, specifically in transformers, we obtain the best results. The height at which the curves begin is noteworthy. This is due to the fact

that we have masked frames that are not evaluated, which give a result of 0, as well as totally correct predictions.

Noise Modeling

In this section, we evaluate the performance of the models in terms of noise level. First, as we previously described in chapter 3, we obtain the noise perturbation of our error distribution for rotation, translation, and scaling, and, finally, we generate a Gaussian noise for the fingers. Finally, we apply a factor of % of the application of this noise.

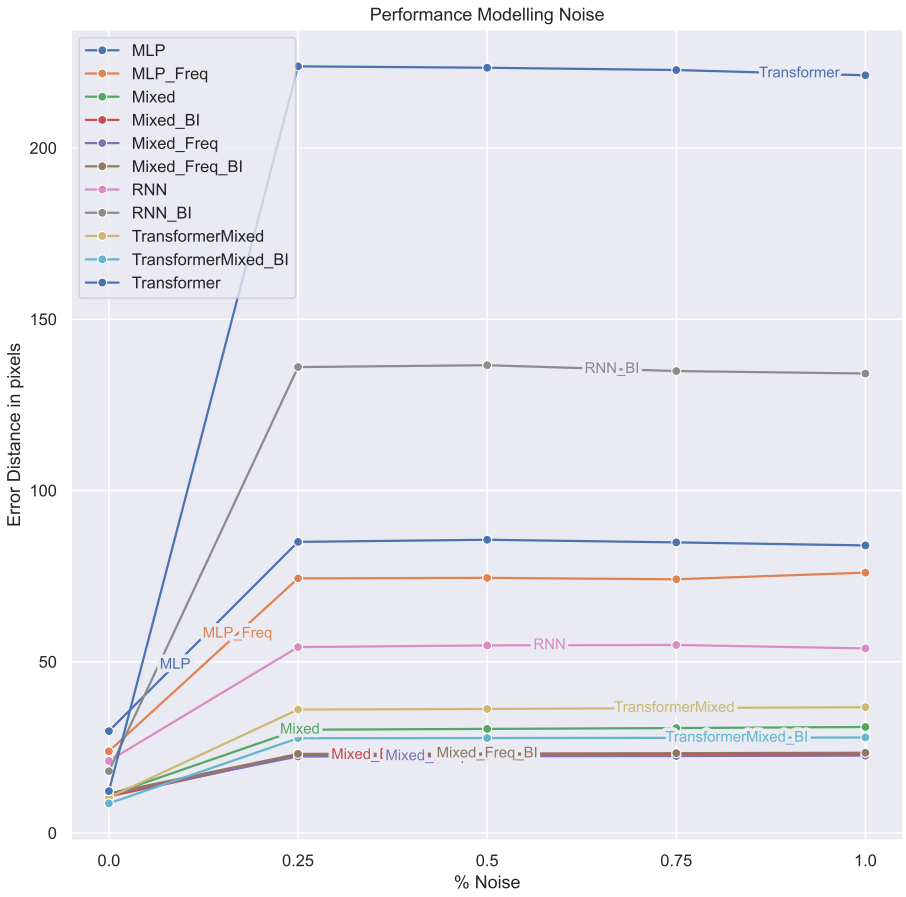

FIGURE 5.3: Performance of the 100 observation length models respect to noise application.

In the following Figure 5.3, the origin of the x-axis (value 0) corresponds to the original data with no perturbations. First of all, we can observe that noise is better tolerated by the frequential model. Second, we find quite striking that the recurrent bidirectional model is more sensitive to noise than the non-bidirectional recurrent model. This could be due to the fact that it is assimilating more noise than it can handle. Finally, we have the transformer giving a worrying result. Compared to the mixed model, it can be seen that once the noise is applied, the temporal transformer

would position the hand in a bad location giving a much larger error distance. On the other hand, mixed models appear to be much more robust to noise, being basically concerned about how to correct the pose, once the trajectory problem is solved.

5.1.5 Visual results

Next, we will present visual results. We will show three of the most pertinent cases of the study. These images will illustrate the following distribution of images in the same plot. First, we will present the original uncorrected data with the corresponding error calculation for the same corrected pose (ground truth). The next set of errors are the MSE in image space, and the MSE (fingers and root-relative) in relative space and Euclidean distance. We will observe our ground truth from the corrected data in the second column. The third column shows both poses, the corrected pose in blue and the ground truth in red to see the differences between poses. Finally we will see output of the algorithm along with their respective metrics with respect to the ground truth.

First Case: Not a good enough estimation.

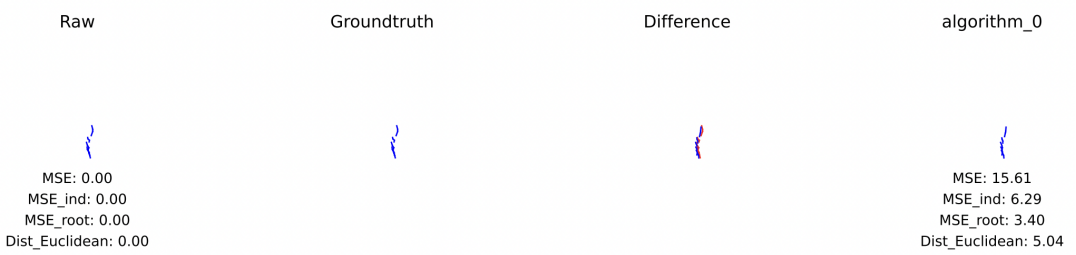

FIGURE 5.4: Raw: Original noisy estimation, Groundtruth: Intepolated Frame, Difference: Ground truth (Red) + Denoising Model Output (Blue), algorithm_0: Denoising Model Output.

Image Figure 5.4 illustrates the problem we encounter with clean data that has not been interpolated. Metrics (MSEs, Euclidean Distance) are zeros since the Raw and Ground truth are identical poses. Our results are quite good in this instance, however our algorithm should estimate the same pose and not try to correct it so that we have not minimal errors that taint the metrics used in the previous analysis Table 5.1. The poses that are correct should be preserved, so we cannot use these models in production at the moment due to this critical error.

Second Case: Corrupt segments

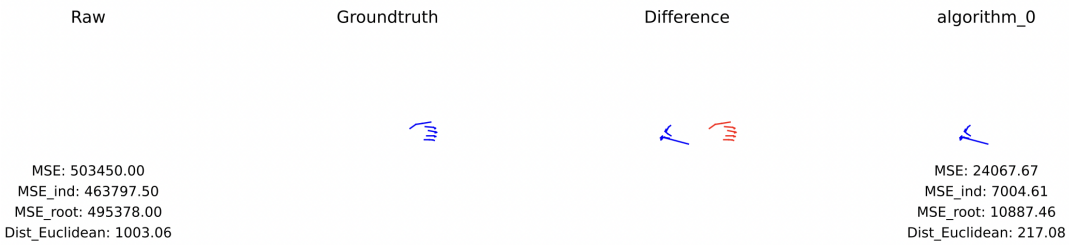

FIGURE 5.5: Raw: Original noisy estimation, Groundtruth: Intepolated Frame, Difference: Ground truth (Red) + Denoising Model Output (Blue), algorithm_0: Denoising Model Output.

In our experiments, Figure 5.5 represents the worst case scenario. The entire observation window is corrupted, so there is no information available. As a result, our models are attempting to predict a pose for which no past or future data is available. The consequence is that our models predict a completely random and misshaped hand.

Third Case: Improving results

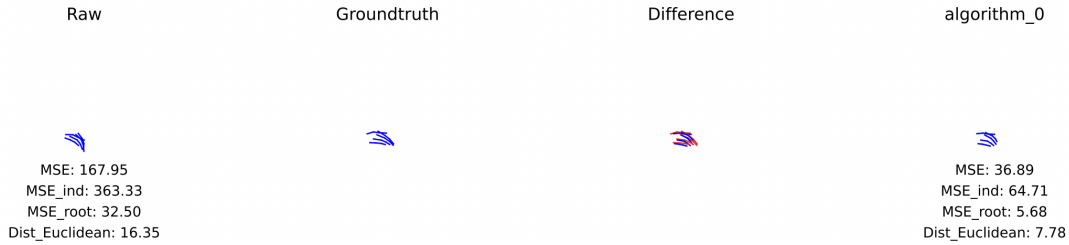

FIGURE 5.6: Raw: Original noisy estimation, Groundtruth: Intepolated Frame, Difference: Ground truth (Red) + Denoising Model Output (Blue), algorithm_0: Denoising Model Output.

The image above Figure 5.6 shows that the results have been improved as compared to the original data, resulting in a significant reduction in error. Nevertheless, it is not the ideal solution as in these cases the metric would be close to zero respect to ground truth. We can see in the plot that the trajectory (where the hand is placed) is relatively perfect and that it is the closest value to 0 in relative space; however, although the exact pose is quite similar, the significant error is related to the hand position. There is a slight improvement over the average of all the poses in terms of the Euclidean distance, because this distance is less than the global distance shown in the Table 5.1 table.

5.1.6 Discussion

In our multiple experiments, we obtained very interesting findings related to the behavior of the models. Furthermore, we addressed the problem of trajectory in an efficient manner, even beating the overall result in the original data. In this section, we discuss all these points.

In the first place, only 1 experiments out of 11 models actually reduces the overall mean error of the original noisy annotations. Pose denoising is a complex issue that requires powerful and complex models.

First of all, we worked with a very small amount of data to address the problem. By rotating the left hand, we show an effective technique to improve our results. We do not really have a ground truth for all the frames, but rather we use the ones that have been fixed. By training on this subset of noisy and correct hand poses, we expected our models to generalize and learn to denoise hand poses under a diverse set of circumstances. Thus, we have little data compared to the total dataset. The big problem we have is when there is no information in the entire observation segment, resulting in noise in our training and in our results, as the algorithms attempt to predict correct poses when there is insufficient and highly noisy data. In addition, inference can follow the coherence of the hand in partially corrupted segments in terms of trajectory prediction, due to the context that gives us the observation window.

Regarding the models, it is important to note that we can now use models that are able to learn from both the past and the future historical context. On the side of using frequential over non-frequential, we see little improvement. The models with the best results, however, are those that are able to handle two different tasks, namely fixing the pose and positioning the hand correctly. In other words, it is still a major problem in the state of the art to address trajectory denoising; even though we are trying to do so with a mixed solution, some papers in the state of the art do not address trajectory and only deal with pose correction.

There is a very prominent problem when predicting poses that are already correct because our algorithms are very rarely able to do so without adding error. In other words, we want our algorithms to predict the same input pose if that is correct. Therefore, it is not yet possible to launch these models in production environments, as we would fix poses but add noise to already correct poses, so it remains a challenge to address this problem.

We were able to overcome the overall error of correct frames counting both the frames that should not be corrected and the frames that should indeed be corrected separately with the use of temporal transformers with recurrents in the transformer decoder and the trajectory decoder separately. The result is positive (less average error), which is a good result and raises the horizons for future work.

Chapter 6

Conclusions

This thesis presents the first comprehensive comparison of state-of-the-art approaches to hand pose denoising. Our best used approach incorporates trajectory and pose as two separate tasks, in contrast to other approaches used in the state of the art, which only address one of the problems, the trajectory problem or the pose problem respectively, and not both at the same time. Several adaptations of recurrent models have been presented using transformers, sequence-to-sequence models and mixed models with recurrent layers in the trajectory as well as other types of architectures for pose denoising. We demonstrate the feasibility of solving hand trajectory problems, and find promising results on hand pose, particularly when using the temporal transformer together with recurrent networks. Mixed models have been demonstrated to be able to absorb and correct for pose noise. Furthermore, we were able to test the hot topic, the use of frequential space. However, we were unable to achieve great results in this regard, but we will continue to explore its use in future work. Finally, our experiments allowed us to identify, analyze, and discuss the key challenges associated with UDIVA v0.5 dataset for hand pose removal, by using an innovative method to generate perturbations in the training data.

Future Work

As a result of this thesis, we have identified many interesting lines of research, which we will leave for future research. It is essential to examine in detail what the frequency models predict, analyzing the frame before the input to the model and the output, in order to determine if the model is really eliminating high frequencies, such as jittering. On the other hand, mixed model design on the trajectory part, it would be interesting to see the use of the frequential only on the trajectory replacing the recurrent as shown in some works allowing to save training and inference time.

Our high priority is to address the issue on frames that do not need to be corrected, where the model should predict the exact same pose. This would greatly reduce the error values in our metrics, and allow to use such models in production environments.

In addition, to increase performance when fixing large data sets, we plan to predict a sequence of frames rather than just the central frame of the observation window, and provide more context when evaluating segments with partial information. Also, we should accurately analyze the influence of the observation window on our data set and how to deal with the problem of totally corrupted segments.

To conclude, it may prove useful to use powerful models such as the Temporal Convolutional network (TCN) (Lea et al., 2017), Graph Convolutional Neural Networks (Wu et al., 2020) or more advanced models of Transformer (Hassanin et al., 2022) such as excitation to further improve the results by learning from a more global context and allow for a more coherent structure in the poses.

Bibliography

Ahmed, Nasir, T. Raj Natarajan, and K. R. Rao (1974). "Discrete Cosine Transform". In: *IEEE Transactions on Computers* C-23, pp. 90–93.

Andriluka, Mykhaylo et al. (2014). "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ballard, Dana H (1987). "Modular learning in neural networks." In: *Aaai*. Vol. 647, pp. 279–284.

Barquero, German et al. (2022a). "Comparison of Spatio-Temporal Models for Human Motion and Pose Forecasting in Face-to-Face Interaction Scenarios". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Vol. 173. Proceedings of Machine Learning Research. PMLR, pp. 107–138.

Barquero, German et al. (2022b). "Didn't see that coming: a survey on non-verbal social human behavior forecasting". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, pp. 139–178.

Carissimi, Nicoló et al. (2018). "Filling the Gaps: Predicting Missing Joints of Human Poses Using Denoising Autoencoders". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Chung, Junyoung et al. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *NIPS 2014 Workshop on Deep Learning, December 2014*.

Cui, Qiongjie et al. (2019). "A Deep Bi-directional Attention Network for Human Motion Recovery". In: *IJCAI*, pp. 701–707.

Falcon, William (2022). *PyTorch Lightning*. Github. URL: https://github.com/Lightning-AI/lightning.

Gauss, Joela F. et al. (2021). "Smoothing Skeleton Avatar Visualizations Using Signal Processing Technology". In: *SN Comput. Sci.* 2, p. 429.

Gu, Chunzhi, Shuofeng Zhao, and Chao Zhang (2022). "Diversity-promoting human motion interpolation via conditional variational auto-encoder". In: *International Workshop on Advanced Imaging Technology (IWAIT) 2022*. Vol. 12177. SPIE, pp. 688–692.

Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos (2018). "Densepose: Dense human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306.

Hassanin, Mohammed et al. (2022). "CrossFormer: Cross Spatio-Temporal Transformer for 3D Human Pose Estimation". In: *ArXiv* abs/2203.13387.

Haykin, Simon (2004). "A comprehensive foundation". In: *Neural networks* 2.2004, p. 41.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9, pp. 1735–1780.

Ionescu, Catalin et al. (2014). "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7, pp. 1325–1339.

Johnson, Sam and Mark Everingham (2010). "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation". In: *Proceedings of the British Machine Vision Conference*.

Katircioglu, Isinsu et al. (2021). "Dyadic Human Motion Prediction". In: *ArXiv* abs/2112.00396.

Kim, Do-Yeop and Ju-Yong Chang (2021). "Attention-Based 3D Human Pose Sequence Refinement Network". In: *Sensors* 21.13.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Lea, Colin S. et al. (2017). "Temporal Convolutional Networks for Action Segmentation and Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012.

Li, Bo et al. (2019a). "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4286.

Li, Jiefeng et al. (2019b). "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10863–10872.

Lin, Jiahao and Gim Hee Lee (2019). "Trajectory Space Factorization for Deep Video-Based 3D Human Pose Estimation". In: *BMVC*.

Lin, Tsung-Yi et al. (2014). "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision (ECCV)*.

Mao, Wei, Miaomiao Liu, and Mathieu Salzmann (2020). "History Repeats Itself: Human Motion Prediction via Motion Attention". In: *ECCV*.

Mao, Wei et al. (2019). "Learning Trajectory Dependencies for Human Motion Prediction". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9488–9496.

Marcard, Timo von et al. (2018). "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". In: *European Conference on Computer Vision (ECCV)*.

Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee (2019). "PoseFix: Model-Agnostic General Human Pose Refinement Network". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7765–7773.

NVIDIA (2021). *CUDA, release: 11.3*. URL: https://developer.nvidia.com/cuda-toolkit.

Palmero, Cristina et al. (2021). "Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset". In: *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1–12.

Palmero, Cristina et al. (2022). "ChaLearn LAP Challenges on Self-Reported Personality Recognition and Non-Verbal Behavior Forecasting During Social Dyadic Interactions: Dataset, Design, and Results". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Vol. 173. Proceedings of Machine Learning Research. PMLR, pp. 4–52.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*, pp. 8024–8035.

Rong, Yu, Takaaki Shiratori, and Hanbyul Joo (2021). "FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1749–1759.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536.

Scarselli, Franco et al. (2009). "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20, pp. 61–80.

Schuster, Mike and Kuldip K. Paliwal (1997). "Bidirectional recurrent neural networks". In: *IEEE Trans. Signal Process.* 45, pp. 2673–2681.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 3104–3112.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Wang, He et al. (2021). "Spatio-Temporal Manifold Learning for Human Motions via Long-Horizon Modeling". In: *IEEE Transactions on Visualization and Computer Graphics* 27, pp. 216–227.

Wu, Zonghan et al. (2020). "Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yadan, Omry (2019). *Hydra - A framework for elegantly configuring complex applications*. Github. URL: https://github.com/facebookresearch/hydra.

Zhang, Song-Hai et al. (2019). "Pose2Seg: Detection Free Human Instance Segmentation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 889–898.

Zheng, Ce et al. (2021). "3D Human Pose Estimation with Spatial and Temporal Transformers". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11636–11645.

Zhou, Kanglei et al. (2021). "STGAE: Spatial-Temporal Graph Auto-Encoder for Hand Motion Denoising". In: *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 41–49.