Dancing2Music In Style: Latent Manipulations in Movement Space

Cem Nurlu [1]

Abstract

Choreography generation is a complex task that is not just a creative one but also a technical one. Like its counterpart music, composing choreographies are a technical process as much as it is a creative process. Composing a choreography requires the choreographer to make certain stylistic choices depending on music, genre and movement. Most of such elements can be identified by eye as these choices exhibit certain rules and constraints on the movement. However, dances are created to accompany the music, which these choices are mostly dependent upon. With recent developments, Generative Adversarial Networks has been a major success in recreating creative tasks that were previously seen human-only. It is also possible to extract, analyze and manipulate information in the latent vectors of such models. While the Dancing2Music paper is great at generating dance sequences from audio, it also maps music to movement in its latent space. In this project, We will explore and analyze this cross-modal latent space of the audio-visual aspect of dance and test if it can be cross-manipulated with success.

1. Introduction

Dance is defined as "the movement of the body in a rhythmic way, usually to music and within a given space, for the purpose of expressing an idea or emotion, releasing energy, or simply taking delight in the movement itself." Simply by its definition, one might believe that dance is purely creative, that it does not have any constraints other than the body itself. However, as in every art medium, dance is grouped and separated into different disciplines known as dance genres. These genres either impose or motivate to move in a certain way, So that they have a certain idea behind the movement. In tap dancing, the footwork will be fast, "tappy" while the body will move freely. While in waltz, the dancers will follow a line, hold their muscles straight and, rise and fall per action.

Choreographing, which is the act of writing a dance piece, just like composing a musical piece. choreographies follow the rules of genre while adhering to the music and the general movement theory. Additionally, the choreographer can choose the style in such the usage of space, energy, emotion, storytelling, aesthetics etc. With so many elements it is no surprise that writing a choreography is a complex task.

To compose a choreography, one must analyze the music carefully. Digital music is stored as wave files. This is a very large file that contains data the size of 10MB+ per minute of song. In real life, one does not care about the wave of the sound, but the elements contained within. The beats, the percussion, chords and progression are what makes a sound a music. Therefore we would need a feature extractor to get rid of the unnecessary parts and turn it into a more densely packed information.

Likewise, analyzing poses directly also is inefficient. To describe a pose, you do not need to know every rotation angle of every limb. Additionally, for a dance video, analyzing every pose in every frame is unnecessary to grasp the choreography, therefore there must lie an understanding behind that connects those movements. The change between the poses are the essential part to understanding the movement. That is why it is necessary to represent both the dance and the music in a way that is more descriptive. In mathematical terms, we need a good dimensionality reduction.

Generative Adverserial Networks (GAN's) have recently shown highly promising results on automating and generating realistic creative works that were previously thought to be only possible for humans. Some of the such networks use encoder-decoder architectures to compress and structure information. Such compressed latent spaces are the spaces that we will be doing our work on. These latent spaces also carry the property that makes vector arithmetic

[1]Department of Physics, University of Boğaziçi. Correspondence to: Cem Nurlu <cem.nurlu@boun.edu.tr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

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

Figure 1. A Neutral Choreography

on them meaningful and possible. However, since this space is fully abstract and complex, basic change of information requires direction finding first. Feature changing via this latent manipulation is a new field and mostly only worked on images as features are more prominent to the eye.

In technical terms, generating dance choreography is a cross-modal sequence generation. Cross-modal audio-visual perception has been a long-lasting topic in psychology and neurology, and various studies have discovered strong correlations in human perception of auditory and visual stimuli. Despite works in computational multimodal modeling, the problem of cross-modal audio-visual generation has not been systematically studied in the literature. Multimodality however, has been extensively studied and researched in text-image modalities. Recent breakthroughs like CLIP uses the idea to match the latent spaces of both modalities into an identical space using contrastive loss for cross-modal generation of images from text or vice versa. In certain ways, changing the features of one of the modalities might come up to be a hard task. Instead, it is possibly easier to manipulate the other matched modality and find the matching latent of the intended modality. With the very recent updates like StyleCLIP, one might use this latent matching to extract the movement of text and change features on the image using these prompts. Since we are going to explore latent spaces of genre, we will need to have the latents of dances from these genres.

The making of audio-visual dance datasets is an extremely tedious and an expensive task, Main challenge coming from the visual part of the pair. There are two ways to create a dataset. The first one is to hire dancers/choreographers to dance to a music in front of a motion capture rig. Datasets created from this require many hours of setup for a small amount of data in exchange for precise data. One might instead use the method of the CLIP, which is to extract data from the world wide web. Many people upload their dance videos to Youtube, which Dancing2Music uses as a baseline for their dataset. Since we do not have access to that dataset, we can use a similar method with a trick. Extracting poses out of a dance video is generally accepted as a hard task due to the fact that pose estimation algorithms are not precise enough. So, there will be missing, wrong and inconsistent frames of poses. Cleaning up or processing this requires manual attention which would take a lot of time. Given that the model of Dancing2Music matches the two latent spaces together, instead we may opt to use the music that was written for that genre. Given that we will only use these to extract the latent of a dance, collecting music for the dataset will be a much easier task than collecting videos and extracting poses out of them to create the dataset.

In this paper we will inspect and analyze this cross modal latent space and the coherence of the cross-modal latent manipulation in audio-visual space.

2. Related Work

2.1. Cross-Modality Generation

This task is built onto a cross-modal generator that is inputted an audio signal and gets motion in return. While the audiovisual space is largely still unexplored, most of the developments in this area was in generation between texts and images. However, audio compared to the text and images is much less organized/structured and more difficult to model the relations. The work that was made on audiovisual mostly lies on speech. These are lip sync models to generate positions of mouth from speech data. There also is audio to image generation by inputting video and trying to reproduce the sounds. The lack of the research in this area comes from also the lack of large datasets for this domain, as there is not much data to work on and the creation of the data in this area is a large hassle.

2.2. Audio To Human Motion Generation Dance Generation

The motion synthesis methods rely on 3 different inputs. The first one is to not use any pose estimation and rely on the network to understand, generate and render realistic videos directly. This method is an inefficient one as one relies on

a single model to generalize to 3 tasks at the same time instead of specializing different models for different tasks. The results here are mixed, as the other results combined with vid2vid.

Previous work in audio-driven human body animation created movement transition graphs timed to the rhythm, or emotion. Recent work has used hidden Markov models , Gaussian processes , recurrent neural networks, or autoregressive encoder-decoder networks to encode music and motion into a low-dimensional latent space, and then decode that latent space to generate dance animation. Others synthesize dance using graph-based frameworks . The primary concept behind these techniques is to build a database of music-motion pairings, then produce basic dance moves synced to the input music by identifying the optimum linking sequences. Chen et al. introduced ChoreoMaster, a choreography-oriented graph-based motion synthesis framework that generates novel dance movements using a unified embedding space for music and dance clips. GANs were used in a similar synthesis-by-analysis learning paradigm, but solely to generate dance in 2D. New ways to include motion variation into the synthesis process have recently been developed .

2.3. Joint representation of Modalities

Multiple works learn cross-modal Vision and language (VL) representations for a variety of tasks, such as language based image retrieval, image captioning, and visual question answering. BERT has been a huge success in the text world. Based on Contrastive Language-Image Pre-training (CLIP) , a recent model learns a multi-modal embedding space that may be used to evaluate the semantic similarity between a text and an image. CLIP was trained on 400 million text-image pairs gathered from a range of publically accessible online sources. CLIP's representations have been demonstrated to be incredibly powerful, allowing for state-of-the-art zero-shot picture classification on a range of datasets. CLIP is a great example for our work as the CLIP model does a similar work on joint representation as this work. CLIP uses the BERT pre-trained model as a natural language processor as this paper first uses audio processing to extract features from the audio to have a more structured understanding of the data.

2.4. Latent Space Manipulation

By encoding data to a higher-level feature space using encoders, we can have a more structured space to help us change the features of a data. For example, changing the hairstyle of a photo requires good photo editing skills and some man-hours to manually patch a new hair to a subject to ensure it looks realistic. By the working principle of GAN's; they already try to optimize the realisticity of the output. So, if we can change the input of the generator that is a latent vector and manipulate it towards a certain hair, it will create us the intended image by itself. Latent Space Manipulation has been predominantly in the area of images and computer vision due to the popularity of this area given that it being a more easily experimentable, useful and understandable space and mostly due to the StyleGAN model and its increased higher level style space. However, there has been no work on latent manipulation. Firstly, there is none because the models that generate dance motion mostly do not use a proper latent space, and the ones used are not disentangled. Secondly, we are not aware of any paper that establishes stylistic expressions for pose space. In image space, there are many features that one might identify as in rotation, transformation, zoom-in, gender, hair color or age. One of the goals of this paper is also to establish certain expressions and analyze the poses; so we do not only make latent space manipulations but also establish the directions that one might be able to move in them.

3. Methodology

Our generation goal is to generate a realistic sequence of poses that are dancing figures which are conditioned on the input music. The Dancing2Music pipeline uses a 'decomposition-to-composition' framework to learn the poses in a curriculum manner. First it learns realistic movements and later condition the movement generator on the input music to match the songs. Since pose space is continuous in the temporal dimension but the task is a sequence generation, the paper introduces what it calls a dance unit abbreviated as DU. Later, they generate realistic movement by a VAE model called $DU - VAE$ and later match latent spaces of the 2 encodes of music and pose in their model to create a joint representation. The pose generation is made in their model $MM - GAN$. After the initial training phase, the actual generation is made in their testing phase network. It encodes the music into a Z_{dan} and uses the Z_{dan} in G_{dan} to create motion recurrently using the last pose as initial recurrently to match the music length. It later modulates the poses to match the beat accordingly.

We have used the Z_{dan} in the framework for our focus on latent manipulation. Readers must note that the space is a simple VAE space and any manipulations will not be as clean as W, $W+$ or S space as there is not enough disentanglement or coverage. The first experiment that we will make will be to explore the latent space in an obvious direction. For this, we have selected 2 distinctly different genres of dance to interpolate choreography.

The first obvious method to use here would be to extract the dance poses from a dance dataset and extrapolate the mean direction from there. However, by joint representation, we might assume that moving in the music space will match

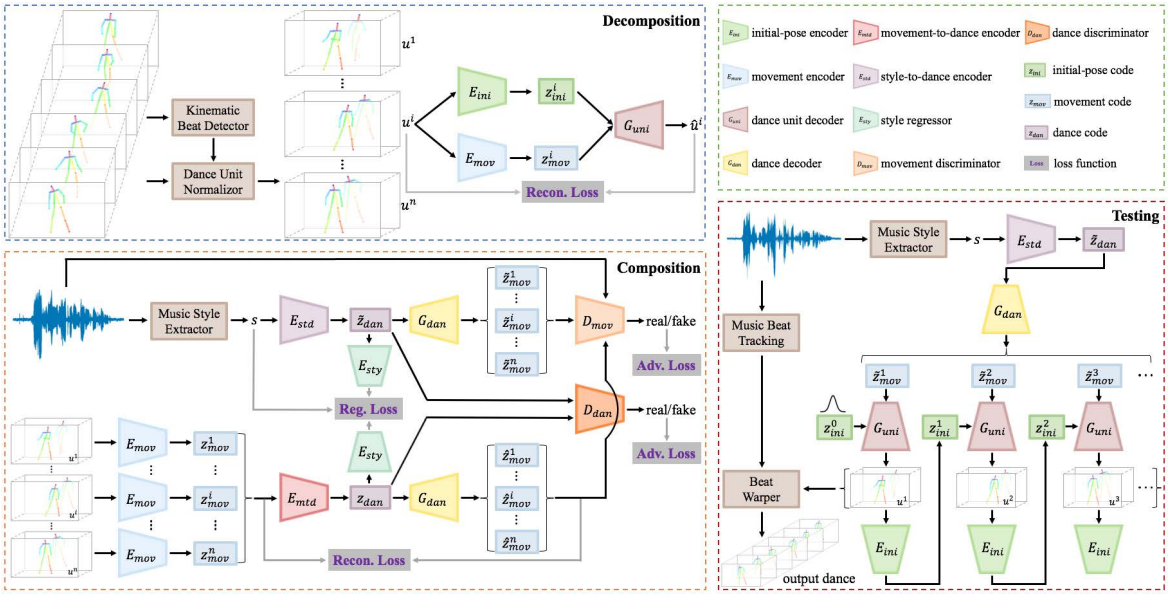

Figure 2. Framework model for the Dancing2Music architecture.

the movement in the dance space. Therefore, instead of a lengthy process of collecting a dataset of dance poses as it also requires an estimation process with non-perfect data, we will be taking a playlist of dance videos in the specified genre and use the music as Z_{dan} generator. Later on we will explore if this choice was a sensible one.

To have an understanding in the latent space, we have created a Dance Music Video Dataset consisting of Ballet and Zumba dance videos. The choice of dance genres were deliberate as they are two contrasting genres. Ballet is considered by dancers as a generally as a more flowing dance. The dancer makes large, smooth movements and the movement is restricted in a disciplined way to not have any unnecessary motions. Zumba is considered by dancers as an energetic dance. The goal is to move as much as possible while still being considered "aesthetic". The dance is characteristically "jumpy" and twitchy. The movements are rough and harsh instead of flowy and slow.

To generate the dataset, we have scraped videos from Youtube. The videos were chosen manually to take the best representations from their genre. In total we have scraped 62 Dance Music videos from various channels, for a total of 6 hours and 35 minutes of dance music. Later, the videos have been converted to .wav format to accommodate the model. The wav files have been preprocessed and trimmed to remove non-music parts and parts where no choreography is available. It was also the authors' goal to add identity loss to the choreographies, but due to the limited amount of data to train an optimizer, it was left out.

Since Dancing2Music was not made for our purposes, we made the necessary changes to the model and the base to allow us to make latent manipulations. We modified the model to allow the extraction and input of the latent vector and also fixed the codebase as the given model did not work as intended in the first place.

Latent space arithmetic is common and easily accepted as sensible in text-image cross modalities. Since to our knowledge there is no previous work on the latents of the pose space, we will also explore the latent space arithmetic in audio/visual space and interpret their success. We will start from a genre and move the choreography to another space to

Since many stylistic features can not be benchmarked in a way to generate a "score", the analysis is made subjectively and its left to the reader to decide if and how much the given examples represent the features. Given that pose space is not "inherent" to a person as much as an image of the face is, the latent directions in these images are also may not be obvious to the naked eye. However, it is commonly accepted between the Image latent manipulation researchers that subjective analysis of images is to be made, so we will be making many subjective analysis here. For the inexperienced reader to the dance domain, we will try to explain the reasoning behind our analysis as much as possible.

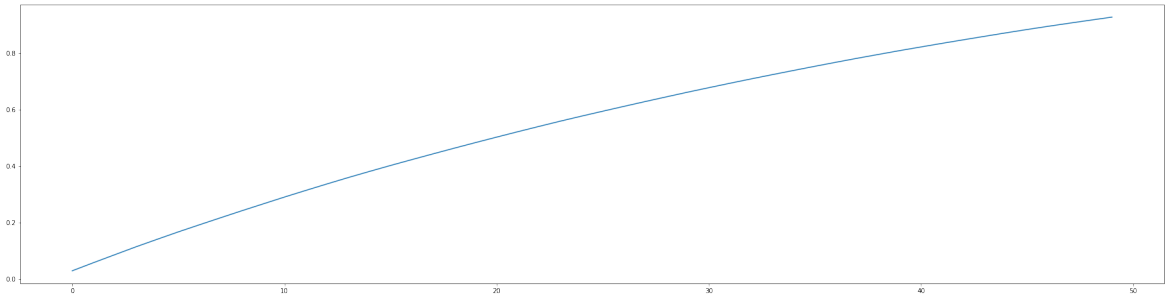

Figure 3. Cumulative sum of Nth Component PCA's explained variation.

4. Experiments

The first experiment made was to see if latent space arithmetic was possible. The goal is to do:

$$Zumba\ Choreography - Mean\ Zumba + Mean\ Ballet = Ballet\ Choreography$$

We used the testing phase of the Dancing2Music Framework and extracted the latent spaces. To match the sample size of each genre, 3 hours of audio per genre is used. Later on we used the testing phase of the Dancing2Music Framework to get the latent vector encodings of each music. The music was taken as a whole to create an accurate Z_{dan}. After the encoding, the latent vectors of each sample are taken and separated by their genre. Each vector created is a 512 length vector. We then take the mean of each genre's latent vector to get Z_{ballet} and Z_{zumba}.

We first made PCA and T-SNE to our data points to see if there was a visible distinction in the dimensionally reduced latent vectors. However the explained variation in the cumulative sum of the pca components suggests that the vectors can not be reduced to

We used the created zumba dances and removed the average zumba latent vector. We call this first interpolation a "neutral" choreography. After the creation of the neutral choreography, we have added the average ballet latent vector to the neutral choreography. This will be the interpolated ballet choreography. We have provided in the appendix link to a non-cherry picked interpolation since does not allow videos.

4.1. Analysis of Results

We will be analyzing the results from per example basis. Since most of the comments made here can be generalized to other examples, we will use the appendix results as our basis. Our talking point will be 3 analyses that are based on the: Original choreography, Neutral Choreography, and the Ballet Choreography in the order of interpolations. To remark once more, this will be a fully subjective analysis as there is no objective benchmark for such a task. The latter analysis is in the eyes of a dancer instead of an ML Researcher.

4.1.1. Original / Zumba Choreographies and The Mean Zumba Choreography

The original zumba choreographies have a diverse amount of difference between them. However while they have different poses and endings, they share the similar characteristic of being an energetic look to them. Since it is a dance created as an aerobic exercise, it has an abundance of movement and the original choreographies have a lot of translational movement and raising and lowering of hands. Figure shows how energetic the dance is by the kinetic energy calculated from each frame:

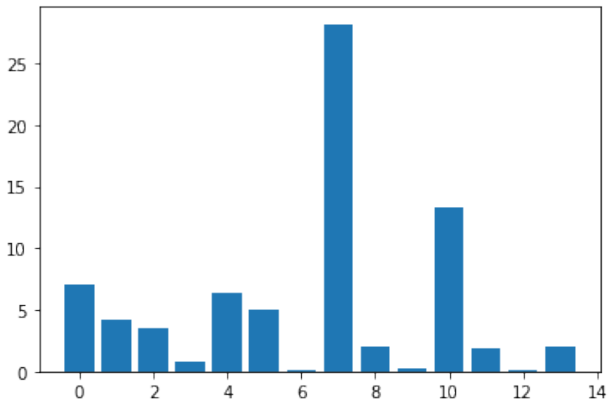

Figure 4. Kinetic Energy of a Zumba Choreography. The x-index represents the limbs in CoCo Format. It can be seen that Lower Arm movements are predominant in the dance.

Figure 5. Visualization of 2 Component PCA analysis and 2 Component PCA+T-SNE. Label 0 represents Zumba and Label 1 represents Ballet

4.1.2. Neutral Choreographies

Neutral Choreographies that are generated from the extraction of zumba mean latent vectors are an interesting one. They are not jumpy and don't have swinging movement as much as the zumba ones do but still are much faster and jerkier than to be considered ballet choreographies. The interesting part is the movements are similar to a genre called the contemporary dance with its philosophy being free from genres of dance and being a more movement focused free dance. So the similarity to this genre implicates us to a certain degree it is somewhat successful in being a "Neutral" dance.

4.1.3. Zumba to Ballet Choreographies

Zumba to ballet choreographies are generated by the given latent vector arithmetic. Ballet choreographies that are generated this way shows the characteristics of ballet. There are poses that were normally jumps becoming a leap in style of ballet. However, the choreographic elements other than genre tend to move away from the original one. While some of the translational motion and rotational movement and the sequence being similar to the original, it is by no means the same choreography still. This indicates that we needed a latent optimizer with a similarity loss to get rid of this issue.

5. Limitations

The main limitation of this study comes from the GAN framework of Dancing2Music. The Latent of the model is not disentangled nor exactly jointly represents the modalities. The lack of a proper feature extractor model and instead the use of the MFCC limits its ability to understand music and the beat since it extracts these manually. This paper could be improved with a better baseline model.

Other than that, all the analysis made other than the dimensionality of the latent is fully subjective and another reader/expert may debate that the analysis might have a confirmation bias attached to them. We believe that the bias might have affected our analysis as it was not a blind critic.

The dataset that dancing2music is trained on is created only from 3 different genres. So it does not have access to a proper understanding of dance but more generates moves in the mean of these 3 dances. This is currently acceptable as there is a lack of dataset in this area but the KL Loss does not exactly represent the results and they will definitely get better with more diversity added.

Lastly, The manipulations made in this paper are not precise. There is no similarity loss for poses and a quantitative approach to understand the poses or a model to discriminate the features. So an optimizer was not possible. In future work, finding quantizations of such features will be a major step to make the interpolations more smooth and the manipulations closer to the original one.

6. Social Impact

This paper introduces a small dataset created from YouTube, and generates dance choreographies according to the manipulations. Further, it could have been argued that creating a genre-agnostic framework to create choreographies would have been a helpful tool to a choreographer to use as a baseline and inspiration for their own work. However, this method is in its too early steps to be taken literally as a dancer to use outside of research. The created dances are not fully coherent nor good enough aesthetically to consider as an actual dance.

7. Conclusion

With this paper, we show that it is possible to find interpretable latent directions in the pose space/ dance space.

This interpretable directions while with this paper manually discriminated, with the use of a proper discriminator model, more features can get available throughout the research of this problem.

Also, one important thing that this paper shows is that latent manipulations of a joint representation of latent vectors in audiovisual cross-modality is transferable. By making changes in the latent space of the music, we can achieve changes in the pose space. This can open up more ideas as we can manipulate music to change its beat, elements add/extract instruments more easily than changing the angles of a pose. We believe that future work in this area is highly open and it is a fresh problem space with useful outcomes.

A. Links

Google drive link to non-cherry picked (randomly selected) dance sequences: Example Generated Videos.