Grenze International Journal of Engineering and Technology, June Issue

Workflow Analysis and Improvements in Music Driven Dance Moves Generation

B.R Sinchana[1], Aditya Sundar Raj[2], Alan S Paul[3], Basavaprabhu G Kiragi[4] and Surabhi Narayan[5]
[1-5]PES University, Department of Computer Science, Bangalore, India
Email: sinchu2304@gmail.com, {sundarrajaditya,alanpaul303,basavaprabhu1652002}@gmail.com, surabhinarayan@pes.edu

***Abstract*— The broad and dynamic fluidity of human motion has always been difficult to understand from the view of computations. Decoding and improving on such a field is of great importance for research and the automation industry. Dance is the culmination of human art, culture and evolution. The expressiveness of the motion is a field of interest and generating such realistic motion through computer intelligent models is a great way of understanding the intricacies of human motion generation. Through this research project, we would like to propose improvements in the workflow of motion generation with just an audio input. The subjectivity of the art form enables us to further explore the workflow both quantitatively and qualitatively. Dance diversity has been a constant point of interest due to the possibilities of interpretation and subjectivity among humans. We focus on exploring a new approach of preprocessing audio files of varying tempo using "jukemirlib" and then continue with the "actor critic reward function" to include as much diversity in the generated output through our work all the while remaining in the context of music to dance with much emphasis given to visual evaluation.**

***Index Terms*— Dance Generation, Motion Generation, AI Choreography, Dance Diversity, Music to Dance and 3D dance model.**

I. INTRODUCTION

Human art forms have been significant in shaping the course of cultural evolution. One such art form that enables total freedom of bodily expression and supplies a joyous sense is dance. Dance as an art form drastically varies depending on the place, the time and the people involved. Concisely, it is heavily subjective and requires a holistic understanding of the intricacies involved to judge the rhythmical patterns or the aesthetics of it. Understanding human motion and tracking movement has been a task that is difficult to achieve in complete reality even with the present advancement of technology. The uncertainty in the process has made research and development a challenging task. Dance is a combination of all aspects of human motion and hence, it becomes a canvas for newer ideas and improvements. The subjectivity of different dance styles also ensures that quantitative and qualitative measures are considered. Previous work in the domain and around the specific problem has made remarkable progress. Our limited resource workaround due to the complexity of the task has led to a few important results and conclusions. It is also important to note that when dealing with motion generation, the real world uses are endless due to the vast applications that will arise from studies in the domain. Through our work, we introduce a way to remove machine specific dependencies in the working process due to the usage of cloud enabled services to achieve

Grenze ID: 01.GIJET.10.2.532

results in a restricted but resource light manner. With major focus being on evaluation of musical pieces that are not a part of the training dataset, suitable evaluation is required. Through the combination of better feature extraction and a "reward system based fine tuning" [22], we have generated both acceptable and plausible dance at the expense of "long term dependency" [24]. The "AIST++" [2] dataset is the pivotal resource for human motion and dance generation due to its comprehensive collection of annotated videos capturing diverse dance choreography. With meticulous annotations detailing human poses with required contextual information, the dataset becomes the richest source for our research. Its significance lies in its ability to facilitate the development and evaluation of algorithms and models for action recognition, pose estimation and pose manipulation. This emphasis makes it particularly valuable for applications in human motion synthesis and dance generation.
The inclusion of local or regional dance forms and genres is highly complex given the particular absence of human motion dance capture in a useful environment. The non uniform tempo of the wild input music pieces and the length of the input plays an integral part in generating the 2D and 3D dance poses. This requires multiple resizing steps of the intermediate data to suit the final output that can be "rigged onto a 3D character" [3]. Since dance is often interactive and depends on the environment that the dancers create around them, all studies on the subject are limited to derivation from existing data which happens to be purely focusing on dance movements.

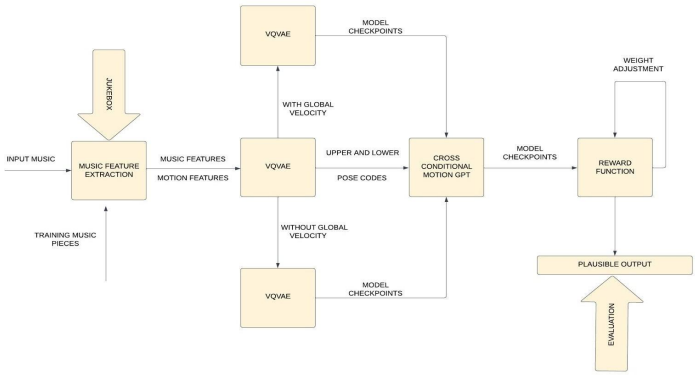

Fig 1. Pipeline of the model

All the extensive research done on motion generation has proved that there is massive room for improvement but the subjective aspect of the results make it very interesting for further inference. Data availability, costs and resource constraints lead to such limitations.

II. Related Work

Profound studies have been carried out using various techniques in order to enable acceptable dance generation. The following studies have set the baseline for this problem with the usage of varying models and technologies available at the time of research:
"Genre-Conditioned Long-Term 3D Dance Generation Driven by Music", a study by Yuhang Huang et al. [11] proposes a new system of "generating dance sequences based on the genre of the input music". This study introduces the inclusion of different classes of music and the corresponding dance sequences for the same. The use of "Long Short- Term Memory" [23] techniques helps in generation of longer sequences of dance moves but there is inherent motion freezing due to the cyclic carry forward nature of "Recurrent Neural Networks" [17]. The data set is a collection of "2600 motion sequences classified by 10 genres of music" and was made publicly available for licensed research. The authors have proposed a set of evaluation criteria that includes "measures of realism and synchronization".
"Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning", a study by Ruozi Huang et al. [10], proposes an improvement for the "genre conditioned neural network" model. Here, "curriculum learning" [4] is introduced to deal with the errors that would be carried forward in the normal working of "Recurrent Neural Networks" [17]. Minor increase in complexity with each training phase will ensure that the errors are minimized. A new data set called the "Dance Generation" data set that consists of "32600 videos of various dancers" is used to train the model. Better evaluation is achieved as a "Dance to Music" data set is used to test the performance of the model. The proposed method achieves "91% accuracy in matching the beat and rhythm of music".

"AI Choreographer: Music Conditioned 3D Dance Generation with AIST++", a study by Ruilong Li et al. [13], proposes a "novel multi modal dataset called AIST++" [2] which includes "over 100 hours of dance" and a combination of "multiple motion capture views" of dance performed by multiple choreographers in multiple styles. Major importance is given to data collection and preparation thus making "AIST++" the "richest source of data on human dance and music". With the new data set, a novel "Full Attention Cross Modal Transformer" is proposed which uses the "AIST++" data set with a "music encoder and a seed motion dance decoder" to generate dance sequences that are comparable to the ground truth. Limited user studies and output diversity make it important for the inclusion of more evaluation metrics.
"Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory", a study by Li Siyao et al. [22], proposes a "novel Generative Pre-trained Transformer model" that makes use of a reinforcement learning method called "Actor-Critic" [9] and a way to remember the weights using a "quantized code book" called "Choreographic Memory". This workflow also makes use of the "AIST++" [2] dataset with more focus given to increasing dance diversity by dividing the motion body into "upper and lower halves". The motion split enables better synchronization, and the reward system reduces erroneous motion generation. Achieving human-like "Temporal and Spatial Coherence" in the output sequence is highlighted as it has the "best quantitative performance out of all proposed models". Fewer user studies and training data dependency reduces qualitative accuracy.
"EDGE: Editable Dance Generation From Music", a study by Jonathan Tseng et al. [24], proposes a "diffusion-based model" that generates realistic and long-form dance sequences conditioned on music. The model is evaluated on multiple automated metrics and a large user study to find that it achieves state-of-the-art results on the "AIST++" [2] dataset and generalizes well to in-the-wild music inputs. It also emphasizes the ability of the model to demonstrate powerful editing capabilities, allowing users to freely specify both "temporal and joint-wise constraints". The proposed method is able to create arbitrarily long dance sequences via the "chaining of locally consistent", shorter clips while it cannot generate dance sequences with very-long-term dependencies.
All the mentioned approaches have proved to show acceptable results with their data and the model used. In our work we will specifically work with the following:

a) 2D and 3D key points related to motion capture from "AIST++" dance database [2].
b) Musical pieces with tempo varying from 80 to 130 to cover a wide range of music from "AIST++" dance database [2].
c) "jukemirlib" [12], an extractor for "OpenAI's Jukebox" [9] for audio preprocessing.
d) A pipeline of pre-trained "transformers" and "auto-encoders" [22].
e) "Reward based reinforcement learning" for fine-tuning [22].
f) Mapping key points to a mesh, here "SMPL" [14].
g) Rigging the "fbx" file using a character from "Mixamo" [18] for better visualization.
h) Visual evaluation of results with scope for quantitative evaluation.

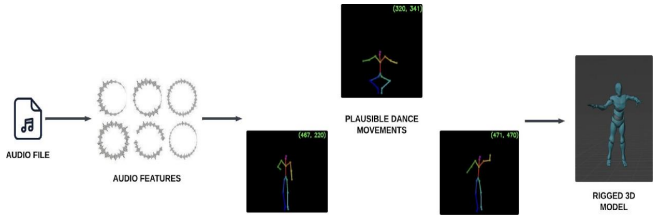

Fig 2. Real time workflow for dance generation

III. Proposed Approach

A. Dataset Description

All the data that is used in our work is derived from the "AIST++ Dance Motion Dataset"[2]. It is constructed from the "AIST Dance Video DB" [1] with multiple choreography styles used for each and every single musical piece. "9 cameras" are set up for the best possible motion capture. "60 musical pieces" with tempo varying from "80 to 130" are used. Every dance genre has "6 musical pieces". The dance genres are further grouped as "New School" and "Old School" based on the dance origin. The weights and key points are available for direct use. The raw video footage is also made available for specific processing tasks.

B. Audio Preprocessing

“OpenAI’s Jukebox” [8] has been of tremendous use in our preprocessing. Given an artist, a style and some lyrics, “Jukebox” [8] can create a new and unique song right from scratch. The method of discarding irrelevant information from music and restricting it to a lower dimension is used. Compression followed by generation ensures useful results. “jukemirlib” [12], a python library, is used for quick and easy extraction of representations from “Jukebox” into the “VRAM” which perfectly suits an environment like “Google Colaboratory” [5].

C. Execution Environment

All our work is done on “Google Colab” due to the lack of physical “GPUs”. The “Colab Pro” membership gives access to a powerful “V100 GPU” and higher amount of “RAM” for in-memory generation. For persistence, we mounted our “Google Drive” to the notebook. All write instructions are guaranteed for an active runtime but there is a delay due to the vast amount of data being generated.

D. Audio Features Analysis

To start with, all the necessary audio files are subjected to preprocessing. “librosa” [16] and “jukemirlib” [12] is used in order to extract the following audio features:

a) “mfcc”
b) “mfcc delta”
c) “audio percussive”
d) “chroma cqt harmonic”
e) “onset env”
f) “tempogram”
g) “beats feature”
h) “acoustic feature”

It is generated twice considering the down sampling, “60 fps and 7.5 fps”. Now, the additional usage of “jukemirlib” [12] issues the removal of unnecessary audio characteristics and reduces the dimensionality.

B. Model Pipeline and Components

For the generated dance to be meaningful, it is required that quantization of poses is done instead of direct audio to motion mapping. A “VQ-VAE” [25] is used to incorporate quantization into the architecture. It is a type of generative model that learns a “probabilistic mapping between input data and a latent space” consisting of an “encoder”, a “decoder”, and a “probabilistic latent space”. Discrete and quantized vectors will lead to more structured and interpretable latent spaces when compared to continuous values. These “quantized vectors” are obtained by assigning the encoder’s output to the “closest vector” in a “predefined code-book” [22]. The “code-book” serves the purpose of grouping similar vectors in the latent space.

A reference point is necessary in vector computation, The root, which is the representation of the total system, will be useful in the context of motion generation. For a skeletal mesh, the most complete global point would be the hip joint representation. Global velocity with the hip joint as the root is computed for relative positions of the rest of the joints. Due to the presence of a root joint, there can be a clear division of planes to increase the possibilities of motion generation. As proposed in “Bailando” [22], the division of the mesh into upper and lower planes will increase the diversity of the generated dance poses due to the greater possibilities of combination. The use of separate “decoders” is important for the generation of consequent frames for each half of the skeleton. Next, the frequency of evaluation is important to adjust the weights of the model since the “checkpoints” [19] will be used further in the pipeline. A “Cross Conditional Motion GPT " is required next in order to keep the “plausibility of the upper and lower body motions” in check. The dual conditioning for better synchronization between motion explains the cross conditional nature of the model. “Bailando” [22] has proposed a very powerful “GPT” that shows how a “Cross Conditional Attention” layer is better than both “casual” and “Full Attention layers” for the music, upper and lower body components.

Due to the sequential nature of human motion generation, repeated intercommunication of the 3 components is required for carrying the term dependency. It is designed in such a way that the generated information is only carried forward but never backwards. “Supervised training” with “cross entropy loss” is used as the learning parameter for the model.The supervised training leads to a less flexible and restricted output with “low coherence” of audio features such as the “tempo”. This calls for an additional step involving a “reward function” for acceptable poses to be given more weightage. “Actor critic” [15] is a “reinforcement learning” architecture that is a combination of “value based” and “policy based” learning methods. Due to the continuous nature of the information space, it is the best learning method. The “actor” is responsible for “selecting actions”. It defines a “policy”, which is a mapping from states to a “probability distribution” over actions. The goal of the actor is to learn a policy that maximizes the expected “cumulative reward” [15].

The “critic” component evaluates the actions chosen by the “actor”. The “critic” helps the “actor” by providing feedback on the quality of the chosen actions. The “critic” maintains a “value function” which estimates the expected “cumulative reward” starting from state ‘s’ and taking action ‘a’, or just the value of being in state‘ s’.
The reward functions are as follows:
a) “Beat-align reward” [22].
b) “Half-body consistency reward” [22].
The “beat alignment reward” is used to evaluate the generated motion with respect to the audio input. It penalizes the absence of a dance beat for the interval that has a music beat. The “Half-body consistency reward” is computed so as to enhance the “synchronous combinations” and avoid “non-plausible combinations” of the skeletal halves. These reward functions ensure that good results are given more weightage and carried forward while the bad results are not carried forward. The “supervised learning” parameter is now changed to an “on-policy decision model”.
The “pose codes” that are generated need to be visualized for better acceptance due to which images and videos from these images need to be generated. The information is also dumped into a “Python pickle file” on further joint configuration using the “SMPL” [14] connotations. “3D model rigs” can be used with suitable “bone retargeting” on “Blender” [6] for the “best rendered” animation.

IV. Experiments

The execution and testing environment is as follows:
a) Hosted “Colab Pro” Notebook.
b) Single”Nvidia V100 GPU”.
c) Colab “High RAM”.
d) Mounted “Google Drive”.
e) “PyTorch 2.0” [20]
To compare the results of our approach and the existing approach, pre trained weights are used to avoid retraining of the model. “Colab” offers only a single GPU for training purposes so the use of a workload manager for multi training is not possible. The “VQ-VAE” [22] has checkpoints [19] generated “without global velocity” followed by the “inclusion of global velocity”. The “cross conditional motion GPT” makes use of the “VQ-VAE” [22] checkpoints to generate full fledged dance poses that are “plausible” and “coherent”. The results need to be fine tuned using the “actor critic policy maker” [22]. On fine tuning, it is observed that the positive rewarded pose codes are “coherent(temporal, spatial)” and “plausible”. “Quantitative metrics” were calculated on the results due to the availability of ground truth motion. All results were achieved as mentioned in “Bailando” [22]. The 3D “SMPL pickle files” that will be used to convert into “fbx” files have not been made public. To test our approach, the 60 original music pieces used by “AIST++ dance database” [2] were preprocessed using “jukemirlib” [12] for better feature extraction. The “VQ-VAE” [22] was trained with these musical pieces and the corresponding motion key points first without global velocity and then with global velocity. The generated checkpoints are then used by the “Cross Conditional Motion GPT” to generate dance poses. Only the required dance pose “json” files, “pkl” files and the checkpoints were “written to drive” due to the heavy write overhead and delay in writing thousands of image frames to drive. This is sufficient for the final output since the images and videos are majorly for visualization purposes. The “actor critic” component gives visually comparable and faster results than what is achieved with “Bailando” [22]. The limitation with the output is that the “long term dependency” cannot be maintained due to the “feature extraction” of audio that picks out the “best length of uniform tempo” in an otherwise “nonuniform musical piece”. It is important to note that the dance at time ’t’ seconds and the dance at time ’t+1’ need to be both plausible and smooth. There should be very minimal or “zero motion freezing”. For “wild” musical input pieces, the results that can be observed with fine tuning in “Bailando” [22] are “qualitatively” acceptable. The results converge after “30 epochs” of training. Due to the absence of “ground truth” for these musical pieces, it becomes important to include a quantitative measure. “EDGE” [24] proposes a measure that is only dependent on the output called “Physical Foot Contact” [24]. This measure can enable the conclusion of the goodness of the results even with the lack of raw ground truth data when visual evaluation proves inadequate.

V. Results And Discussions

20 “wild” input music pieces were preprocessed. Since “lyrical context” is not taken into consideration, we have used instrumentals of popular songs which have been trimmed to maintain an “average tempo range of 80-130” and lengths between “28 seconds to 46 seconds”.

For complete model testing, we have included songs that are usually associated with the dance genres as follows:
A) “New School”
B) “Old School”
“New School” consists of “Ballet Jazz”(“Bird Set Free”, “Dandelions”), “House”(“Hotel Room Service”, “Passionfruit”), “Krump”(“Little Weapon”, “Multiply”), “LA Hip Hop”(“Heroes”, “Target Practice”), “Middle Hip Hop”(“Flashing Lights”, “Heaven’s EP”) and “Street Jazz”(“Levitating”, “On the floor”).
“Old School” consists of “Break”(“Juice”, “Knock you out”), “Lock”(“Get Lucky”, “Uptown Funk”), “Pop”(“Moth to a flame”, “Stay”) and “Waack”(“This is what you came for”, “Timber”).
Comparing the results obtained on wild input for, it can be observed that with “uniform tempo” and “short term dependency”, our approach is better than the results obtained through simple “actor critic fine tuning”. The “physical plausibility” can be measured using the “Physical Foot Contact (PFC) score” [24] on the “pkl” files that are generated with suitable formatting. Due to the difference in mesh features and “SMPL” [14] joint configurations, the calculation parameters of the results are dependent on the preprocessing used. The formatting of the output to suit required physical measures such as “Physical Foot Contact” [24] are not dealt with here but can be used for better proof of acceptance. In an unbiased poll on the premise of an in-person demonstration with over “500 attendees”, “68%” of votes favored our output with the conclusion that it is visually better structured.

TABLE I. COMPARISON OF THE AVERAGE TIME TAKEN TO SET UP THE DANCE CONTEXT

Method of preprocessing	**Average time taken for smooth transition**
Method A(without “ jukemirlib”)	8.8 seconds
Method B(with “jukemirlib”)	3.1 seconds

VI. CONCLUSION AND FUTURE WORK

While the obtained results are only acceptable, it can be made better by improving the dependency between the generated dance for a whole musical piece instead of specific parts thus ensuring “long term dependency”. A common confusion is the genres of dance and the genres of music. In common practice, the music may belong to a particular genre but the dance that is choreographed for it can belong to a different genre. Likewise, for a particular audio piece belonging to “genre A”, the generated dance may or may not belong to “genre A”. Classifying the output dance based on genre would be very useful for choreographers who may be facing a “creative block” by providing more clarity and options to choose from. The context of lyrics and inclusion of “facial expressions” is a complex task since raw data needs to be captured for the same but it is definitely achievable. “SMPL-H” provides skeletal meshes with facial points that can be used to our advantage. We can also make use of the “OpenPose” [7][21][26] library to work with small and specific sets of dance styles or dance videos that are not available as part of the “AIST++” [2] database. This ensures better diversity and variety through the inclusion of more genres with their corresponding choreography in the field of human dance generation. Whole-body (Body, Foot, Face, and Hands) 2D “Pose Estimation” and Whole-body (Body, Foot, Face, and Hands) 3D “Pose Reconstruction” can be achieved with this method.

REFERENCES

[1] AIST Dance Video Database. https : / / aistdancedb.ongaaccel.jp/.
[2] AIST++ Dance Motion Dataset. https://google.github.io/aistplusplus dataset/.
[3] Ilya Baran and Jovan Popovic´. “Automatic rigging and animation of 3d characters”. In: ACM Transactions on graphics (TOG) 26.3 (2007), 72–es.
[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. “Curriculum learning”. In: Proceedings of the 26th annual international conference on machine learning. 2009, pp. 41–48.
[5] Ekaba Bisong. “Google colaboratory”. In: Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners (2019), pp. 59–64.
[6] Blender. https:// projects. blender. org/ blender/ blender. git/.
[7] Z.Cao, G.Hidalgo, T.Simon, S. Wei and Y. Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[8] Prafulla Dhariwal et al. “Jukebox: A generative model for music”. In: arXiv preprint arXiv:2005.00341 (2020).

[9] I. Grondman, L. Busoniu, G. A. D. Lopes and R. Babuska. “A survey of actor-critic reinforce- ment learning: Standard and natural policy gradients”. In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012), pp. 1291–1307.
[10] Ruozi Huang et al. “Dance revolution: Long-term dance generation with music via curriculum learning”. In: arXiv preprint arXiv:2006.06119 (2020).
[11] Yuhang Huang et al. “Genre-conditioned long-term 3d dance generation driven by music”. In: ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2022, pp. 4858–4862.
[12] jukemir lib. https : / / github . com / rodrigo - castellon / jukemirlib/.
[13] R. Li, S. Yang, D. A. Ross and A. Kanazawa. “Ai choreographer: Music conditioned 3d dance generation with aist++”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 13401–13412.
[14] M Loper, N Mahmood, J Romero, G Pons-Moll, M.J Black. “SMPL: A skinned multi-person linear model”. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023, pp. 851–866.
[15] Maja J Mataric. “Reward functions for accelerated learning”. In: Machine learning proceedings 1994. Elsevier, 1994, pp. 181–189.
[16] Brian McFee et al. “librosa: Audio and Music Signal Analysis in Python”. In: Jan. 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003.
[17] Larry R Medsker and LC Jain. “Recurrent neural net- works”. In: Design and Applications 5.64-67 (2001),p. 2.
[18] Mixamo. https://www.mixamo.com/.
[19] Ron A Oldfield et al. “Modeling the impact of check- points on next-generation systems”. In: 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007). IEEE. 2007, pp. 30–46.
[20] Adam Paszke et al. “Pytorch: An imperative style, high- performance deep learning library”. In: Advances in neural information processing systems 32 (2019).
[21] Tomas Simon and Hanbyul Joo and Iain Matthews and Yaser Sheikh. “Hand Keypoint Detection in Single Images using Multiview Bootstrapping”. In: CVPR. 2017.
[22] Li Siyao et al. “Bailando: 3d dance generation by actor- critic gpt with choreographic memory”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 11050–11059.
[23] Ralf C Staudemeyer and Eric Rothstein Morris. “Under- standing LSTM–a tutorial into long short-term memory recurrent neural networks”. In: arXiv preprint arXiv:1909.09586 (2019).
[24] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. “EDGE: Editable Dance Generation From Music”. In: 2022.
[25] Aaron Van Den Oord, Oriol Vinyals and Koray Kavukcuoglu. “Neural discrete representation learning”. In: Advances in neural information processing systems 30 (2017).
[26] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh. “Convolutional pose machines”. In: CVPR. 2016