Applied Research of an End-to-End Human Keypoint Detection Network with Figure Ice Skating as Application Scope

Nadin-Katrin Apel
Stuttgart Media University
apel@nadin-katrin.de

Supervised by
Prof. Dr. J. MAUCHER; Prof. Dr. S. RADICKE
maucher@hdm-stuttgart.de; radicke@hdm-stuttgart.de

Abstract

Human joint detection is a key component for machines to understand physical human actions and behaviors. Especially in figure ice skating, this understanding is an indispensability. There are many difficult figures and poses, even difficult to clearly understand for the professionalized jury. This thesis presents an end-to-end approach to detect the 2D poses of a person in images and videos. The underlying architecture combines three modules: image segmentation, body part recognition, and human joint detection. The applied research reveals multiple findings regarding different training settings, optimizing functions, and motion capturing techniques for the special application scope of figure ice skating. In particular, multiple elaborated concepts are meant to further spur research of pose recognition in artistic sports.

1. Introduction

Human 2d pose estimation has gained more and more attraction in recent years. For example, *Facebook*, one of the *Big Five* technology companies, has published 73 research paper targeting the problem of pose estimation in the last three years. The most popular ones are *DensePose, VideoPose3d* or *Mask R-CNN* [1, 10, 11, 19]. A company in Canada *wrnch.ai* even specialized on keypoint recognition from image and video data with a lot of product options [38]. Furthermore, many enterprises are becoming more and more interested in Sports Content Analysis (SPA) e.g. *Bloomberg, SAP*, and *Panasonic*, just naming a few [9, 24].

But how did this topic get into such a demanded focal point? One of the reasons are the various application areas to which pose estimation can be applied to. Main fields are sports, visual surveillance, autonomous driving, entertainment, health care and robotics [18, 26, 42]. For example *Vaak*, a Japanese startup, developed a software, which would detect shoplifters, even before they were able to remove items from a store. This yielded in a drastic reduction of stealing crimes in stores [30].

The exercise of sport not via visiting a sports course, gym or club became of fundamental severance in 2020, when the Coronavirus SARS-CoV-2 spread the entire world [22]. Many courses such as Yoga, Pilates, or general fitness routines went online and were often conducted via Zoom, Instagram Live, or other video streaming technologies [2]. However, what participants were often missing, was the feedback of the coach on how the exercise was going, and whether it was done right or wrong. So in 2020 more than ever was missed a technology, which is good at pose estimation, or even further, action recognition, in sports.

Most investigations in this field target everyday activities not including complex poses, which can be encountered in professional sports. This is why these architectures often fail when applied to more complex movements. For competitive sports, there are various metrics of high interest depending on the environment. Competition and training can be differentiated as can be sports executed by multiple athletes versus single combats. Team sports, such as basketball or soccer, are interested in predictions about how the other team behaves during the game and which would be the best reaction to their behavior for winning the game. During practice, 2d pose recognition can help to optimize the sports-persons movements by taking the role of a coach. This could provide an answer to the question on how certain activities might be optimized. Single competitive sports with very complex movement routines are for example gymnastics and figure ice skating. Both sports include various artistic body movements, which are not part of daily activities. Even famous and well-rated 2d pose recognition networks such as *OpenPose* or *VideoPose3d* fail to recognize these poses.

1.1. Special Application Scope of Figure Ice Skating

The support of an accurate 2d pose recognition module as coach or jury would make a huge contribution to figure ice skating. For one it has a very complex rating system

which makes the highly professionalized jury a necessity for competitions. Additionally, there exists a shortage of judges and people often complain scoring would not be executed fairly [15, 31, 32].

This is why the here presented thesis investigates 2d pose recognition with a special focus on figure ice skating.

1.2. Applied Research Work

In this work the performance of *VideoPose3d, OpenPose,* and *wrnch.ai* on figure ice skating elements was tested, but the result showed a lot of failed frames, especially spins, such as the artistic Biellmann spin. This rose motivation to further look into datasets, dataset creation and suitable neural network architectures for labeling figure ice skating videos. While at the same time keeping in mind a good performing architecture, which would be able to run on mobile phones.

This is why the creation of a figure ice skating dataset with the help of the XSens motion capturing data in a real figure ice skating arena environment with an example setup in Blender and the help of Makehuman is elaborated in this thesis.

An end-to-end fully convolutional architecture consisting of three modules for background extraction, body part and keypoint detection was developed. Moreover, multiple experiments with network architectures, learning optimization techniques, and loss functions were conducted. As an outlook, the here summarized thesis developed several concepts, where these studies should continue to.

2. Related Work

2d pose estimation sets the baseline for machines to understand actions. It is the targeting application scenario of localizing human joints or keypoints in images and videos. Many research studies explored and researched this topic already with the most popular ones being *OpenPose* and *VideoPose3d* [4, 19].

For 2d pose recognition there are mainly two general procedures: either *top-down* or *bottom-up*. *Top-down* first detects a person and then finds their keypoints. Whereas *bottom-up* first detects all keypoints in the image and then refers to the corresponding people. For top-down it is argued, that if a person is not detected via a bounding-box or alike, no keypoints can be found. This would lead to more undetected keypoints in the frames of a video. When there are many people in the image with many occlusions, people often can not be detected. However, when a person is correctly detected, it is said that accuracy would be higher [17].

A famous *top-down* approach for example is *Mask R-CNN* developed by the *Facebook AI Research* team. It consists of three branches and two stages. The first stage presents the *Region Proposal Network (RPN)*. It proposes candidate bounding-boxes for objects. The second stage performs classification and bounding-box regression by extracting features using region of interest pooling, which they refer to as *RoiPool*. Additionally, *Mask R-CNN* predicts a binary mask for each *ROI* in the second stage. They receive top results in the *COCO* challenges for instance segmentation, bounding-box object detection, and person keypoint detection [10]. Other famous *top-down* approaches include *Simple Baselines*, *the Cascaded Pyramid Network* or *Deep High-Resolution Learning* [6, 27, 39].

One of the most discussed and popular *bottom-up* approaches as of today is *OpenPose*. Their neural net predicts vector fields for the joint connections, which they call *Part Affinity Fields (PAF)*. Additionally, it estimates candidate keypoint locations via Gaussian distributions. These they refer to as *Part Confidence Maps*. From these detections, they refer to the associated human poses. *OpenPose* shows very good results on the MPII and COCO challenges. They highlight their performance, which especially shows it's strength when detecting multiple people. The performance wouldn't change even if more and more people enter the scene. With sufficient hardware equipment, this would even show decent results in realtime [4].

In a newer research, they additionally pay attention to the temporal characteristic in video sequences in their work of *Spatio Temporal Fields*. Their approach is able to track multiple people's poses across frames being runtime-invariant to the number of people in the frames. Besides, they receive highly competitive results on the *PoseTrack* challenges [21]. Some other famous *bottom-up* approaches include *Convolutional Pose Machines* and *PifPaf* [13, 37].

The COCO, MPII, and PoseTrack challenges lead to several studies in the pose estimation field. Nevertheless, their dataset targets rather simple daily activities. There have been only conducted a few studies on competitive sports such as basketball, ice-hockey or swimming [8, 17, 35]. For sports including full-body flexibility or special unconventional jump or turn rotations as can be found in dance, gymnastics or figure ice skating, there have been only a few studies [5, 14, 40, 41].

Studies on figure ice skating topics encounter three main problems: The first is domain knowledge. In C. Xu et Al.'s research [40], they try to predict the technical and performance scores from video data with only the ice skating program as video input and the judge scores as labels. This will very likely not result in useful results since first, the figure ice skating judging system adjusts every year, second, there are always different judges on the competitions who all have their own rating style, and third however, the skater is one of the always winning ones, this skater can do a good program falling at the main elements and still score very well, because the jury tends to be biased. This is one of the main controversies in the figure ice skating fairness of program judgment. Another problem is the missing dataset. There

didn't exist a dataset with joint labeling until FSD-10 [14], which only came out at the beginning of the year 2020. One very interesting study from Yu, Ri et Al. [39] tries to create simulations from the figure ice skating elements. They exactly encountered the problem, that pose estimation currently does not work on difficult pose sequences such as spins or very flexible positions. However, they were able to successfully predict jumps and simple steps from videos into 3d simulation.

3. Dataset

To develop a working machine learning algorithm, the correct dataset according to the problem must be chosen wisely. Since there was no dataset publicly released targeting human joint recognition in figure ice skating at the writing time of this thesis, a proof of concept was developed, investigating how such a dataset could be generated. Therefore, motion capturing data was recorded at the ice rink with the *MVN Link* product from *XSens* [16].

3.1. Inertial Motion Capture XSens (MVN Link)

Since motion capturing has to deal with several difficulties in the ice rink, such as lightning, other skaters, expensive location, the inertial motion capturing technique by XSens showed it's advantages in such a setting. 17 IMU motion trackers were applied to the athletes. They were connected to each other via cables and a battery was attached at the skaters back. Then the skaters could move freely and even move away from the connected router, which transmitted the data to the accompanying software. The data was transmitted later when the skater returned to the hub. However the initial set-up process took about an hour, including the attachment of the trackers and the repetitive calibration process, the tracked data was recorded very well. Another disadvantage was that the cables, trackers, and battery constrained the skater's movements. Since falls are common in figure ice skating this could result in injuries, which is why the athletes skated very carefully. A great advantage was that the recordings could be made during usual skating sessions with multiple skaters on the ice.

3.2. Dataset creation with Blender and Makehuman

The recorded data then was preprocessed and applied to 3d scenes in the open-source 3d computer animation software Blender [3]. Moreover, with a prototypical approach, the animation data was transmitted into a valuable dataset applicable for machine learning algorithms of pose recognition tasks.

4. Method

For the here presented end-to-end keypoint recognition architecture three modules were built to extract the background, find the body parts in the image and detect the human joints. However, even if the background extraction was not important for other keypoint recognition architectures such as *OpenPose* [4] or *VideoPose3D* [19], it is important in here, since this recognition architecture should be able to be used during practice when there are multiple skaters on the ice, but only the focused skater should be recognized. With the body part detection module, the well-established approach from *OpenPose* was altered in a way to not recognize vector fields, which the joints connect, but recognize the visible body parts. For the keypoint recognition module, a Gaussian with a radius of three pixels and a standard deviation of three was calculated. In Figure 2 a demonstration shows an example frame labeled by these three modules, showing Alena Kostornaia, the 2020 European champion during her program.

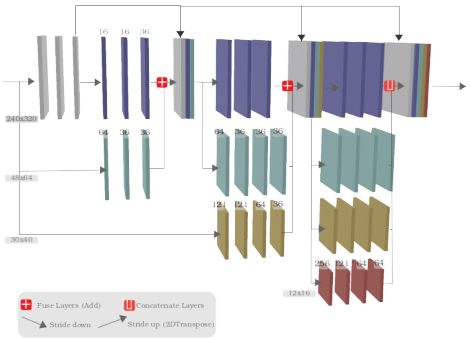

Figure 1. HRNetV3: Developed High-to-low resolution network architecture

All in all, a fully convolutional architecture was built with three networks, that all are based on high-to-low representation learning. This architecture consists of one input block $\mathcal{N}_I$ and three subsequent blocks $\mathcal{N}_L, \mathcal{N}_M, \mathcal{N}_S$ and $\mathcal{N}_{XS}$. These subsequent blocks combine feature maps with lower coarser representations with the original sized input image feature maps and thus learn the features of the different levels equal to the HRNet strategy [12, 36].

However, different from the *HRNetV1* and *HRNetV2* pooling is not used to decrease or increase the size of the feature maps. These changes of receptive fields are conducted with usual convolutions and associate strides $\mathtt{s}$ and kernel sizes $\mathtt{k}$, with $\mathtt{s} = \mathtt{k}$. To increase the feature maps a transposed convolution is applied with again $\mathtt{s} = \mathtt{k}$ according to the strided down convolutions. This allows the network to learn additional weights for the upward and downward convolutions and improve these level exchange processes.

Furthermore, the layer blocks in the first and second stage are fused to combine the mentioned feature levels. In the third stage, however, all feature levels are concatenated to fully exploit the multi-resolution convolutions as argued

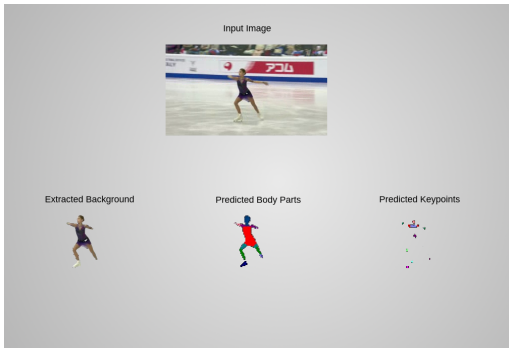

Figure 2. Learned labels by the three modules: extracted background, human part detection and keypoint detection *skater: Alena Kostornaia, 2020 European champion[33]*

in *HRNetV2* [12].

As visualized in Figure 1 the number of feature maps was adjusted for the different blocks. Moreover, in the first stage, only three convolutional layers were used for one block, in the other stages four layers where used for the lower levels and only in $\mathcal{N}_L$ three convolutional layers were applied for the blocks throughout the network.

Another adjustment is, that the input image as initial input was added to all block levels but the $\mathcal{N}_{XS}$ block, which uses the fused layers of all the preceding levels as input. Every last block of the stages combines the information from all previous block levels and additionally the input block $\mathcal{N}_I$ was added.

Every block is completed with batch normalization and a *selu* activation function. The output of the network is predicted by a linear softmax activation function. For the background-extraction and keypoint detection network *mean-squared-error* calculates the loss, and in the human part detection network a custom loss function was developed, to optimize the network's weights. In sum, the presented network comprises 3,008,562 parameters of which 3,003,930 can be learned. The amount of all layers for the network is 156.

5. Human Parts Module: Developed HRNetV3

In table 1 the in this research work developed network architectures are outlined. These range from 88 until 177 layers and are all fully convolutional networks. For the HRNet *(traditional)* the HRNetV2 presented in [12] was slightly deviated by replacing the *pooling* layers with *strided-down* and *transposed convolutional* layers. The number of levels and sizes of the feature maps correspond to HRNetv3. As in HRNetV2 4 stages are used with a filter size of 64 for each stage. One additional level per stage starting with just the highest level $\mathcal{N}_L$ and the concatenations of the levels from stage 2 ongoing are demonstrated with HRNetV2. For the HRNet *(filters)*, the filters are adjusted in a way, so

Table 1. Ablation Human Parts Module: Network Architecture Comparison

Name	Parameter Amount	Trainable Parameters	Layers	Time / Epoch
HRNet *- traditional*	5,595,221	5,589,237	171	84.76s
HRNet *- adjusted filter*	4,936,997	4,933,029	171	66.23s
HRNet *- 3 stages*	848,409	846,441	100	82.33s
HRNet *- stride-down-up*	4,953,269	4,949,157	177	63.21s
HRNet *- stride-down-up-input*	4,185,605	4,181,493	177	80.57s
HRNet *- add-input*	4,185,605	4,181,493	177	80.57s
HRNet *- add-depthwise-conv*	3,182,342	3,177,654	156	83.10s
UNet	656,389	654,261	88	85.37s
HRNet *- v3*	3,008,562	3,003,930	156	85.43s

that higher levels use fewer filters than lower levels. Additionally, the filter amount of the *convolutional* layers is decreased in the level blocks, starting from the highest filter amount, and reducing until 36. All levels are *concatenated* with a filter size of 36. Subsequent architectures all make use of the adjusted filter amount. For the HRNet *(3 stages)*, only the three first stages from *HRNet - filter* are used and $\mathcal{N}_{XS}$ is omitted.
In HRNet *(stide-down-up)*, the Input from 240x320x3 is strided down to 120x160x36 and this layer is used as *Input* layer and highest level block $\mathcal{N}_L$. Lower layers scale relatively to $\mathcal{N}_L$. The HRNet *stride-down-up-input* uses the initially strided down *Input* layer as input for every stage instead of the concatenated results of lower levels. As presented in the *MobileNet* paper [25] in *HRNet - add-depthwise-conv depth-wise convolution* are experimented with and the *Concatenation* layers are replaced after each stage with *Add* layers. In the UNet all levels are used and combined via *concatenation* according to the traditional UNet [23]. Finally, the resulting high-to-low resolution network HRNet *(v3)* is presented, where the first stages are fused and only the last stage is concatenated. Furthermore, some adjustments to the layer amount in the levels and stages were applied as presented in Figure 1.

5.1. Network Comparison

All network architectures trained comparably fast.They mainly showed different results on different dates, depended on other tasks running on the server.

The U-Net 1 has the smallest amount of layers with 88 and 654,261 trainable parameters. The HRNet *stride-down-up* composes most trainable parameters with 4,949,157 and 177 layers, since the Input layer is first strided down and at the end transposed up again. On the other hand, due to

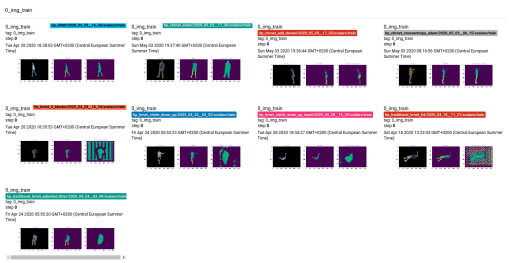

Figure 3. Predicted images of network architectures after first episode

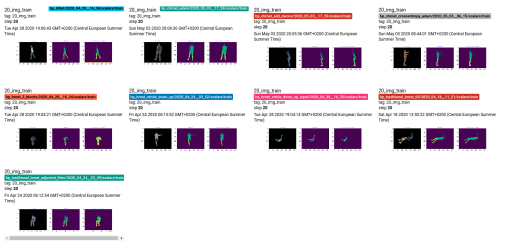

Figure 4. Predicted images of network architectures after 20th episode

Figure 5. Ablation Human Parts Module: Accuracy Comparison

the reduced feature map sizes, this reduces the computation effort as well. Regarding accuracy 5, all comparison figures show very similar curves. These are very steep until the 500th episode and start to flatten then. When looking at 3 and 4 this steep initial increase is visually reflected. Even if, after the first episode the circumferences of the body shapes are predicted vaguely, after the 20th episode the networks have already learned to predict the locations of body parts relatively close to the real ones. Of course, there has to be regarded, that the different predictions were made on different poses, which might differ in their complexity. Nevertheless, the UNet 1 and the HRNetV3 seem to accomplish superior performance. After the first episode, the body shape is already estimated precisely close to the input image. HRNetV3 even shows correct class estimations for different body parts such as the torso, head, arms, and legs already. In the 20th episode then the labels seem to be estimated very close to the true labels.

The accuracy graph shows that the HRNet *(traditional)* and HRNetV3 get the best results from epoch 500 onwards. At step 55.55k they reach 99 percent. However, the HRNet *(stride-down-up)* has a lower curve until the 4.3kth epoch, it then makes a jump in accuracy and at the end reaches 99 percent as well. Interestingly the U-Net with the by far

Table 2. Ablation Keypoint Detection Module: Network Architecture Comparison

Name	Parameter Amount	Trainable Parameters	Layers	Training Time/Epoch
KPS-Dense - *block_L*	221,748,579	218,735,769	51	68.75s
KPS-Dense	23,097,980	20,086,328	13	67.05s
KPS-FCN	4,119,529	1,109,991	41	63.78s

fewest parameters and layers does receive very good results as well making almost 99 percent in accuracy at the end. The HRNet with only three stages shows the lowest curve and reaches only 98.87 percent points on accuracy.

6. Ablation Study: Keypoint Module

For the keypoint detection module experiments with dense layers *(KPS-Dense)* and a fully convolutional network layers *(KPS-FCN)* which implements the third stage of the HRNetV3. KPS-Dense 2, as well, utilizes the third stage of the HRNetV3. The output Dense stage uses max pooling to reduce the feature map size from 240x230x9 to 120x160x1, 58x78x32, and then 28x38x64. The resulting *Pooling* layer is flattened and followed by two Dense layers, which are combined with Dropout and batch normalization. The layers include 1024 and 512 units. The first Dense layer is connected to a Softmax activation function, the second to a linear activation function. The model estimates 38 x,y coordinates for the human joint locations. The network's loss function calculates the distance from the predicted keypoint locations to the true locations and optimizes the networks via mean-squared-error.

For the KPS-FCN network for true labels, Gaussian circles were calculated with a radius size of three pixel for the locations of the keypoints. The keypoint classes are combined for equal body locations such as legs and arms, resulting in 11 joint classes. This network predicts 11 different labels for the classes and is further optimized with mean-squared-error. All networks are trained with an Adam optimizer with a learning rate of 0.001 and run for 5556 epochs. The evaluation of the dense networks show, that they were not able to learn the human keypoints. The networks randomly put the keypoints in the center of the image with no relation to the body parts, because the probability that a person was standing in the center of the image and the location of a possible keypoint in the center is higher.

On the other hand, the small KPS-FCN network was able to recognize the human keypoints in relation to the body parts.

7. Ablation Study: Optimizers

Several different optimizers were tested such as SGD, Adam and Nadam. Furthermore, strategies were applied to the SGD optimizer with a learning rate decrease and a

scheduler which would reset the learning rate to the start value, when learning started to get stuck. These experiments revealed that the Nadam optimizer performed best. It started with a steep progress and did not seem to get stuck but continue learning. Adam was very close to these results. However, SGD reached a local plateau earlier, and the network could not efficiently continue with the learning process.

8. Ablation Study: Custom developed loss function CIloss

For the human parts detection module a custom loss function was developed to countervail the problem of class imbalances. The background pixels appear most often, and the different body part classes occur much less often and they even differ a lot in their relative occurrence. For the keypoint and background detection module *Sparse Categorical Cross Entropy* was used.

This problem is confronted with a weighed map μ, which takes the body parts as a graph and calculates the distances from each body part b_x to all other body parts b_n, and stores the data inside a table. The weighed map μ is applied to the true labels of y_{true} so that wrong predictions further away from the true class will be punished more. For example, if the network predicts hand instead of lower arm, the error will be less than if the network would have had predicted foot. Additionally, this weight map is evened out with a multiplier to reduce the distances and facilitate the learning process for the network. As in *MSE* the difference $\mathcal{E}$ between y_t and y_p is calculated, but the result is not squared. Instead, the absolute values are used. The resulting error $\mathcal{E}$ is multiplied with the weighed map receiving δ. To calculate the loss, the sum of the error $\mathcal{E}$ and delta δ are calculated pixel-wise:

$$\mathcal{E} = y_t(x) - y_p(x)$$

$$\delta = \theta \cdot \mu[argmax(y_t)]$$

$$L = \sum_{i=0}^{n} \mathcal{E}_i + \delta_i$$

9. Conclusion

Inspired by popular research from *OpenPose, VideoPose3d* or *Simulation of figure skating* [4, 19, 41] which all have problems to correctly predict human joint locations for spins in figure ice skating, this thesis analyzed several aspects such as motion capture with connection to dataset creation techniques and network architectures with the influence of optimizer and loss functions.

The result of this research work is a successfully created new high resolution fully convolutional neural network HRNetV3, which includes state-of-the-art research findings from I.a. HRNet, HRNetV2, and MobileNet [12, 25, 36]. This algorithm accomplished to predict human joint locations by learning from the synthetic dataset 3DPeople [20]. In sum, three modules for background extraction, human parts detection, and keypoint recognition were created. These modules partly correspond to concepts used in other state-of-the-art research such as *OpenPose*, however, the single ice skating domain specific background extraction module stands out other architectures.

For the applied research, a Python 3.7 project was built with Tensorflow 2 [29] as core library. Several experiments were performed in multiple Docker containers with the Tensorflow:*latest-devel-gpu* as base image [28] In fact, lots of effort was spent to write good readable code with a decent OO-style, following the *SOLID* principle. So the created Github repository, which is planned to be made publicly available, might help to promote further research on this topic. Especially, the in this work used feature-driven approach making use of Github's feature pull request style and the formulation of proper commit messages helped to iteratively improve the presented project architecture in a continuous style.

Since the data is one of the main aspects deciding whether a neural network learns the desired predictions, a look at motion capturing was taken and the generation of an according dataset investigated. Further, an appropriate figure ice skating dataset for human joint recognition did not exist at the writing time of this thesis. To create a prototypical dataset some recordings were captured with the inertial motion capturing set from XSens. The setup felt very time-consuming and the cables all over the body connected to the battery and sensors felt very uncomfortable for the skaters on the ice. Additionally, multiple calibrations had to be conducted, probably because the system was not prepared for the typical gliding movements on the ice surface. Nevertheless, after the imposing capture, the integration into Blender and use of MakeHuman for creating a synthetic dataset became very promising. Especially the diversity of data or video sequences possible from just one motion capture recording is highly valuable. This is why it is very convincing that this is the way to gain decent data for an artistic sport such as figure ice skating. However, due to the limitations of this thesis, a complete dataset could not be created, but research towards applicable ones found *3DPeople*. Yet, critically reviewing *3DPeople*, where random background images were put behind the moving actors, results could improve, if an ice arena simulation was added to Blender as a more realistic scene. Another encountered caveat appeared later when using real video sequences. Regarding spins, they include a huge amount of motion blurriness especially for skaters with high levels such as Olympic or world championships winning Evgenia Medvedeva or

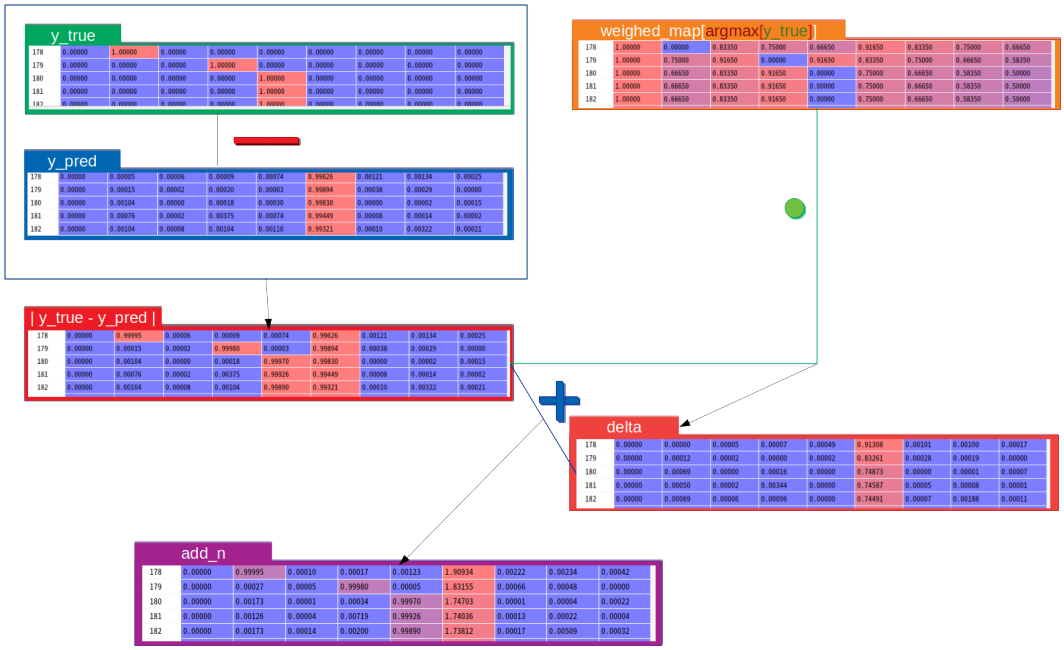

Figure 6. Visualization of the in this thesis custom developed loss function CILoss

Alina Zagitova [7]. Nevertheless, the data from *3DPeople* included only image sequences without any blurriness, being one of the reasons why the algorithm has difficulties to correctly predict human joints for spins such as the Biellmann pirouette.

Throughout this research, further ideas on how to continue on this topic of human joint recognition arose. First, the creation of a decent synthetic dataset for figure ice skating with added motion blurriness and a more realistic background scene for the recordings in Blender. Furthermore, testing other motion capturing methods as for example the *Awinda* set from XSens, which could be much more comfortable and faster to apply, because the sensors are wireless, and no cables restricting movements on the ice, must be dealt with. Additionally, tests could be conducted with a markerless system such as Vicon [34]. Moreover, the network architecture could become more compact and better in performance by applying a grid search with multiple different parameters. In fact, the temporal information from a video could add additional information, allowing to faster and more efficiently predict videos as already argued in [21].

This here presented study tries to further spur development and research in figure ice skating to improve fairness in this sport. For example, in [41], they simulated some figure ice skating figures and translated these to 3d animation. It would be very interesting to see, how skaters were rated, if the technical specialists and judges would see the animation, without knowing, who the skater was, or what the skater looked like. This could be a huge step in fairness. In addition, the rating could be conducted remotely, saving the jury from the cold ice rink and driving efforts.
An even more advanced step would be to replace the jury partly, who are sometimes hard to find for a competition. Another application would be the support during practice, with an included action recommendation component, which could help for instance to improve some jumps.

All in all, the here presented research is meant to serve as a foundation for further investigations towards action recognition in sports, especially artistic ones such as figure ice skating. Indeed, a very up-to-date topic on the writing of this paper are restrictions during the lock-downs of cities due to the COVID-19 (Corona) virus. Highly promising topics include as well physiotherapy or feedback during fitness and dance routines.

Books and Articles

[1] Rĩ za Alp Güler, Natalia Neverova, and Iasonas Kokkinos. “DensePose: Dense Human Pose Estimation In The Wild”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[4] Z. Cao et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–1.

[5] Daniel Castro et al. “Let’s Dance: Learning From Online Dance Videos”. In: (Jan. 2018).

[6] Y. Chen et al. “Cascaded Pyramid Network for Multi-person Pose Estimation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7103–7112.

[8] M. Fani et al. “Hockey Action Recognition via Integrated Stacked Hourglass Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 85–93.

[10] K. He et al. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2980–2988.

[12] Sun ke et al. “High-Resolution Representations for Labeling Pixels and Regions”. In: (Apr. 2019).

[13] S. Kreiss, L. Bertoni, and A. Alahi. “PifPaf: Composite Fields for Human Pose Estimation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 11969–11978.

[14] Shenlan Liu et al. “FSD-10: A Dataset for Competitive Sports Content Analysis”. In: (Feb. 2020).

[17] Takuya Ohashi, Yosuke Ikegami, and Yoshihiko Nakamura. “Synergetic Reconstruction from 2D Pose and 3D Motion for Wide-Space Multi-Person Video Motion Capture in the Wild”. In: (2020).

[18] Paritosh Parmar and Brendan Tran Morris. “Learning to Score Olympic Events”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 76–84.

[19] D. Pavllo et al. “3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 7745–7754.

[20] Albert Pumarola et al. “3DPeople: Modeling the Geometry of Dressed Humans”. In: (Apr. 2019).

[21] Y. Raaj et al. “Efficient Online Multi-Person 2D Pose Tracking With Recurrent Spatio-Temporal Affinity Fields”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4615–4623.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *LNCS* 9351 (Oct. 2015), pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[25] Debjyoti Sinha and Mohamed El-Sharkawy. “Thin MobileNet: An Enhanced MobileNet Architecture”. In: (Oct. 2019), pp. 0280–0285. DOI: 10.1109/UEMCON47517.2019.8993089.

[26] Waqas Sultani, Chen Chen, and Mubarak Shah. “Real-World Anomaly Detection in Surveillance Videos”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 6479–6488.

[27] K. Sun et al. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5686–5696.

[35] B. Victor et al. “Continuous Video to Simple Signals for Swimming Stroke Detection with Convolutional Neural Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 122–131.

[36] J. Wang et al. “Deep High-Resolution Representation Learning for Visual Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1.

[37] S. Wei et al. “Convolutional Pose Machines”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4724–4732.

[38] “wrnchAI, BUILT FOR ALL YOUR HUMAN VISION NEEDS”. In: (2020). URL: https://wrnch.ai/product (visited on 05/12/2020).

[39] B. Xiao, H. Wu, and Y. Wei . “Simple Baselines for Human Pose Estimation and Tracking”. In: *2018 European Conference on Computer Vision (ECCV)* (2018).

[40] C. Xu et al. “Learning to Score Figure Skating Sport Videos”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2019), pp. 1–1.

[41] Ri Yu, Hwangpil Park, and J. Lee. “Figure Skating Simulation from Video”. In: *Computer Graphics Forum* 38 (Oct. 2019), pp. 225–234. DOI: 10.1111/cgf.13831.

[42] Nan Zhao et al. “See your mental state from your walk: Recognizing anxiety and depression through Kinect-recorded gait data”. In: *PLoS ONE* 14 (2019).

Online

[2] Sabrina Barr. *CORONAVIRUS: FROM YOGA TO BARRY’S BOOTCAMP BEST EXERCISE CLASSES ON ZOOM, INSTAGRAM AND YOUTUBE*. 2020. URL: https://www.independent.co.uk/life-style/health-and-families/coronavirus-home-workout-exercise-class-yoga-dance-kids-elderly-joe-wicks-a9421126.html (visited on 05/11/2020).

[3] *Blender.today - daily art & development live streams*. URL: https://www.blender.org/ (visited on 05/17/2020).

[7] Liz Clarke. *Russias Alina Zagitova captures gold in womens free skate*. 2018. URL: https://www.denverpost.com/2018/02/22/alina-zagitova-captures-gold-womens-free-skate-russia/ (visited on 06/07/2020).

[9] David Fox. *Video-based sports analytics system from SAP and Panasonic announced at IBC*. 2014. URL: https://www.svgeurope.org/blog/headlines/video-based-sports-analytics-from-sap-and-panasonic-announced-at-ibc/ (visited on 05/11/2020).

[11] Facebook Artificial Intelligence. *Feacebook publications with topic pose estimation*. 2020. URL: https://ai.facebook.com/results/?q=pose%20estimation&content_types[0]=publication&years[0]=2020&years[1]=2019&years[2]=2018&sort_by=relevance&view=list&page=1 (visited on 05/03/2020).

[15] Borzilleri Meri-Jo. *2014 Winter Olympics - Olympic Figure Skating Controversy: Judging System Is Most to Blame for Uproar*. Feb. 2014. URL: https://bleacherreport.com/articles/1969257-olympic-figure-skating-controversy-judging-system-is-most-to-blame-for-uproar (visited on 06/02/2020).

[16] *MVN Animate*. URL: https://www.xsens.com/products/mvn-animate (visited on 05/17/2020).

[22] RKI. *SARS-CoV-2 Steckbrief zur Coronavirus-Krankheit-2019 (COVID-19)*. 2020. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html (visited on 05/11/2020).

[24] Bloomberg Press Room. *Bloomberg Sports Launches Stats Insights, Sports Analysis Blog*. 2012. URL: https://www.bloomberg.com/company/press/bloomberg-sports-launches-stats-insights-sports-analysis-blog/ (visited on 05/11/2020).

[28] Tensorflow. *dockerhub - tensorflow*. 2020. URL: https://hub.docker.com/r/tensorflow/tensorflow/tags (visited on 05/22/2020).

[29] Tensorflow. *TensorFlow Core*. 2020. URL: https://www.tensorflow.org/overview/ (visited on 05/22/2020).

[30] CFI Tom Meehan. *Using Artificial Intelligence to Catch Shoplifters in the Act*. 2019. URL: https://losspreventionmedia.com/using-artificial-intelligence-to-catch-shoplifters-in-the-act/ (visited on 06/02/2020).

[31] Olessia Tom Meehan. *Expertin bei Eislaufweltmeisterschaften: Drei entscheidende Minuten*. Dec. 2014. URL: https://www.stuttgarter-zeitung.de/inhalt.expertin-bei-eislaufmeisterschaften-drei-entscheidende-minuten.3b370419-3df9-4ea7-81ba-97d64480e78f.html (visited on 06/02/2020).

[33] ISU International Skating Unigion. *ISU European Figure Skating 2020 - Ladies - Result*. 2020. URL: http://www.isuresults.com/results/season1920/ec2020/CAT002RS.htm (visited on 05/22/2020).

[34] Vicon. *Award Winning Motion Capture Systems*. 2020. (Visited on 05/17/2020).