2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Unsupervised Sequence Alignment between Video and Human Center of Pressure

Shiwei Jin[1*], Minh Vo[1*], Chen Du[1], Harinath Garudadri[2], Allison Py[3],
David J. Moore[3], Kristine M. Erlandson[4], Raeanne C. Moore[3], Truong Nguyen[1]

***Abstract*— Center of pressure (COP) estimation with images/videos as input achieves accurate precision with the development of the human skeleton joint extraction tasks. As a supervised learning task, correct labels acquired from COP with regard to the input images/videos are significant. Thus, synchronization between these two different types of sequences is necessary. If these two different modalities are misaligned, the downstream tasks' precision is affected significantly due to the inaccurate labels from the COP sequence. In this paper, we used a synchronized dataset and unsupervised deep learning to train an Alignment Network to align video and COP sequences on another unsynchronized dataset where each sequence starts at a different time and has different frame rates. On the synchronized dataset, the Alignment Network removes 84.4% of temporal offset. On the unsynchronized dataset, we proposed a simple yet effective Differential Network to simulate one practical downstream task. We used the differential Network to estimate the sway level of COP. Results show that this method achieved significant improvement (over 20% improvement on three sway level cases) over the misaligned dataset.**

I. INTRODUCTION

When a human body stands on the ground, it exerts a force equal to the body's weight on the ground. The pressure of this force is distributed over the soles of the feet. The Center of Pressure (COP) is the point where the pressure of the body would be if it was distributed at a single point rather than over the entire soles of the feet. COP is an essential indicator of human body sway, which is measured widely in biomechanics for assessing postural stability, gait control and disease patterns. However, measurement of COP requires specific hardware and expertise which limits the measurement to a clinical/laboratory environment. With the ever expanding use of Deep Neural Networks and high-accuracy learning-based human 2D pose estimation [1], [2], [3], image-based COP estimation achieves over 95% precision [4], [5] in recent years.

The key idea of image-based COP estimation is to predict the mapping from one sequence (images/videos of human movements) to another (COP locations/trajectory). Since this is a supervised learning task, correct labels (COP) with regard to the input (images/videos) is needed. To be specific, the alignment between these two sequences, which are in different data types, is of great importance, as shown in Fig. 1. Ideally, the synchronization is achieved among devices at the hardware level. However, the collected sequences can be misaligned due to a lack of relative temporal information. This misalignment results in the dataset restriction for downstream tasks since the ground truth always has a unstable offset given the input.

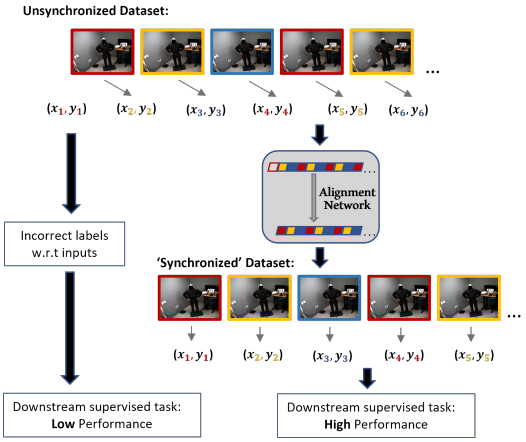


Fig. 1: An overview of the proposed framework. The supervised learning model performance of downstream tasks would be affected due to the wrong labels (caused by the offset between the labels and the input sequences). However, the proposed alignment network can essentially remove this offset, which is beneficial for the downstream tasks.

In order to limit misalignment between different modalities, several synchronized methods have been proposed. Time warping methods, such as dynamic time warping [6] and canonical time warping [7], are traditional ways to align time series because of the changeable-time-axis property. These methods can accommodate missing or added frames, and are applied widely in speech processing. More recently, due to the excellent performance on extracting features from convolutional neural network, the alignment information is estimated based on the concatenated features extracted from different modalities and also achieves high performance [8], [9]. However, these existing alignment methods are mainly applied on the video and audio sequences.

In this work, we propose a neural network to align participant's movement from the corresponding video and COP data similar to the video-audio synchronization. The pipeline is shown in Fig. 1 The specific contributions of this paper are: 1. Proposing a trainable end-to-end Alignment Network that can estimate alignment information between the video and COP sequence using an unsupervised learning

* indicates equal contribution.
[1] Electrical and Computer Engineering Dept, University of California, San Diego, CA 92093.
[2] Qualcomm Institute, University of California, San Diego, CA 92093.
[3] Psychiatry Dept, University of California, San Diego, CA 92093.
[4] University of Colorado-Anschutz, Aurora, CO 80045.

strategy; 2. Achieving high generalization ability across different datasets for the Alignment Network; and 3. Proposing a differential network that can estimate a person's fall risk, which is used to measure the aligned result from the Alignment Network.

II. RELATED WORK

A. Human Pose Estimation

Human pose estimation is an interest-point detection task, enabling machines to find the configurations of the subject's joints and body parts in images and videos. Most approaches [1], [10] utilize a top-down strategy, which is to incorporate a person detector first, followed by estimating the joint locations. Due to the precision of the person detector and various human poses, a recent study proposed a bottom-up approach [11], which was able to detect identity-free body joints for multiple people in an image and correctly linked individual body parts with the correct individual. Based on this idea, OpenPose [2] integrated a learned 2D vector field linking two joints with the grouping method and achieved higher precision. Higher-HRNet [3] combined a joint detection step with coarse-to-fine feature pyramids, thus keeping the image details while also speeding up the joint detection.

B. Modality Alignment

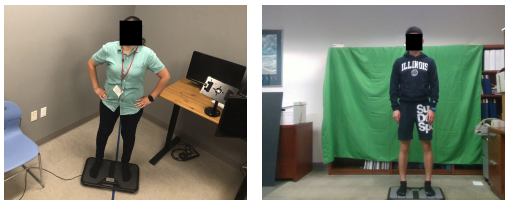
Data collection across multiple devices with different data types occurs frequently in practice. For example, when recording a person talking, we may use a camera to record the visuals and a microphone to record the audio. A common issue arises when these different devices are not properly synchronized: they might have different starting times and might have different sampling rates. As a result, when using data recorded from multiple devices together, the data sequences might not always be synchronized - that is, the n^{th} video frame might not always correlate with the n^{th} audio frame. As a result, sequences of data recorded from the same event would be misaligned if we simply use them frame-by-frame. Thus, adjusting the relative timing among sequences of different data types (modality alignment) is crucial. Automatic modality alignment has been studied over decades in computer vision. Early works [12], [13] synchronized different modalities based on the canonical correlation analysis (CCA). Handcrafted features [14] or learned features [8], [9] from input modalities were utilized in the more recent methods. SyncNet [8] proposed to learn joint embedding features of different modalities and further align them using the dynamic time warping (DTW) method. AlignNet [9] achieved coarse-to-fine alignment based on the feature pyramid structure and replaced the time-warping step with one convolutional layer for extracting more patterns.

C. Video-based Balance Evaluation

Along with the development of human joint detection tasks, several video-based human balance evaluation methods have achieved excellent performance [4], [5]. The basic idea of these methods is estimating the COP or the Base of Support (BoS) from images or videos. In one study, Scott et al. [4] estimated the foot pressure map using the extracted 2D human joints' locations from each frame. In another study, Du et al. [5] used the video sequence to predict the COP trajectory, thus making full use of the temporal information between frames.

III. DATA COLLECTION AND ISSUES

The datasets consist of an unsynchronized dataset and a synchronized dataset.



(a) Unsynchronized dataset (b) Synchronized dataset

Fig. 2: Video dataset samples

A. Unsynchronized Dataset

The participants from the unsynchronized dataset consist of men and women, aged 50-74 and with or without HIV, from UCSD's HIV Neurobehavioral Research Program (HNRP). Participants had varying levels of clinical fall risk (low, moderate, high). Each participant has three videos and balance board recordings.

The balance board records the participant's COP location (x, y) for 20 seconds at 25 frames per second. The camera records the participant standing on the balance board during this period at 60 frames per second. The camera is located above participants and records them at an angle, as depicted in the sample frame in Fig. 2(a). The participants' faces are masked out for privacy purposes. The balance board data sequences and the video frame sequences are not synchronized: we do not know if the video starts recording before or after when the balance board starts recording.

B. Synchronized Dataset

The synchronized dataset consists of seven male participants, students at UCSD and aged 19-23. Each has one video and balance board recording (COP sequence) for a total of seven videos and COP sequences.

The balance board records the participant's center of pressure location (x, y) for 120 seconds at 60 frames per second. The camera records the participant standing on the balance board during this period at 30 frames per second. The camera is located directly in front of the participant and films them upright, as depicted in the sample frame in Fig. 2(b). The participants' faces are not masked during data collection (only here for confidentiality). The COP sequences and the video frame sequences are synchronized: each frame of the video and balance board recording is timestamped, which can be used to align the two data sequences. Thus, the synchronized dataset can be used to train the model to align the video and COP sequences.

C. Issues with the Datasets

In the unsynchronized dataset, the lighting is inconsistent in many videos, making it challenging to use existing image processing techniques (such as optical flow) to find the movement of the participants. Many videos include other people who need to be removed before generating key points using the human pose estimation network. In addition, the camera's locations in the unsynchronized and synchronized datasets are different, so the videos have to be preprocessed using perspective transformation to correct for the camera angles' differences. Lastly, the synchronized dataset is used as the training and validation set since it has timestamps that can be used as ground truths to evaluate the model.

IV. PROPOSED METHOD

Based on the pyramid structure of the AlignNet, which is originally used for video-audio alignment, we propose to modify the audio branch to solve the alignment between the video and COP sequences with the unsupervised learning strategy.

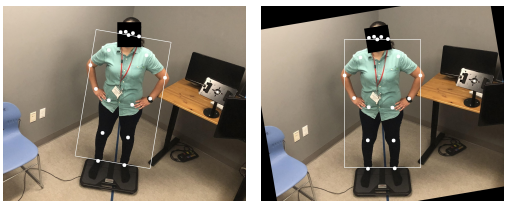
A. Data Preprocessing

All videos are passed into the Higher-HRNet to compute participants' joint locations in the videos. Then, we analyze the joint locations to rectify issues with the dataset, as mentioned in Section III.

(a) Before (b) After

Fig. 3: Removing presence of other people from videos

a) Detecting and Removing Outliers: Given the (x, y) coordinates of joint locations from the human pose estimation network, we use the RANSAC [15] method to detect the presence of joint outliers in videos, which would indicate that there is more than one person in the video. For each video where outlier joint locations are detected, we apply a blur to the left-hand side of the image to remove additional people, as depicted in Fig. 3. Using this method, we were able to remove outliers from all videos.



(a) Before (b) After

Fig. 4: Perspective transform on videos

b) Rectifying Pose: Since the network is trained with the synchronized dataset, we must tilt the participants' pose in the unsynchronized dataset upright, similar to the pose in the synchronized dataset. We do this by first finding the source bounding box around the participant's body. We then find the destination bounding box by making the source bounding box upright and centered in the image. Using the source and destination bounding boxes, we find the perspective transform matrix:

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}$$

Then, using this perspective transform matrix, we find the transformed location (x_d, y_d) of each key joint location (x, y) as follows:

$$(x_d, y_d) = \left(\frac{M_{11}x + M_{12}y + M_{13}}{M_{31}x + M_{32}y + M_{33}}, \frac{M_{21}x + M_{22}y + M_{23}}{M_{31} + M_{32}y + M_{33}} \right)$$

The bounding box and transformation result are shown in Fig. 4. Lastly, we account for the difference in location of the participants in the initial video. We normalize the keypoint locations across all datasets by calculating the average (x, y) coordinate of key joint locations and setting that average as $(0, 0)$. We then recalculate the coordinate of all participants' key joint locations relative to the new origin $(0, 0)$.

B. Alignment Network

The whole pipeline of the alignment network [9] is shown in Fig. 5. The video feature sequence $S^i_{video}, i \in [1, n]$, consisting of the extracted joints locations, and the COP sequence $S^i_{COP}, i \in [1, m]$ are fed into the end-to-end alignment network, which is composed of video and COP branches.

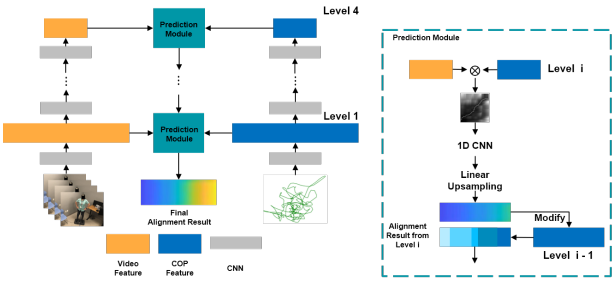


Fig. 5: Structure of the Alignment Network [9]. The left figure is the overall alignment network structure. The video and COP sequences are fed into the network. The alignment result is estimated after feature pyramid extraction and prediction modules. The prediction module structure is elaborated in the right figure.

a) Video Branch: The shape of the video sequence (joint location sequence) is $(17 \times 2, n)$. The first two dimensions are the number of extracted joints and x, y coordinates, respectively. The third dimension is the number of input

frames, which is not fixed. Since different joints play different importance during the COP estimation task, we utilized the joint attention module to learn the weights for each joint. Following the strategy proposed in [16], apart from joint locations, each joint's velocity and acceleration were calculated and concatenated with the original joint location matrix. Then the concatenated matrices were fed into the pyramid structure for fine-to-coarse feature extraction. Each time, the matrices were downsampled by two with average pooling and fed into five 3D convolutional layers for the kth pyramid features, notated as $F^k_{video}, k \in [1, K]$.

b) COP Branch: The shape of the COP location sequence is $(1 \times 2, m)$. We concatenated the corresponding velocity and acceleration matrices with the location matrix. The K-level pyramid features were extracted with the same convolutional layer parameters as those of the video branch instead of the input size. The extracted features from kth level is notated as $F^k_{COP}, k \in [1, K]$.

c) Prediction and Warping Module: Because the same level extracted features from the video and COP have the same shape, we used cosine similarity to model the correlation measurement between F^k_{video} and F^k_{COP}. Then we used two 2D convolutional layers to predict the correspondence from the k-th level correlation map and performed the linear interpolation by two for warping the $(k-1)$-th COP features. The final predicted alignment was available after K iterations.

C. Training and Evaluation Strategy

We used the synchronized dataset to apply the unsupervised learning strategy to train the Alignment Network. Before training, we modified the COP sequence, including 1) randomly deleting the frames, 2) randomly creating the displacement between two sequences, and 3) randomly repeating the frames. The alignment information was also modified accordingly as the ground truth. The loss function is the L2 norm between the prediction from the first pyramid and the ground truth. We used the 85% length of the sequence as the training data. The remaining 15% was used for validation.

a) Evaluation: In order to test the performance of the Alignment Network, we proposed to use both the synchronized and unsynchronized datasets for evaluation. As for the synchronized dataset, the evaluation criterion was followed by [9] in which the Average Frame Error (AFE) was proposed as the average difference between the reconstructed frame indices and the original undistorted frame indices. We applied a leave-one-subject-out protocol where one participant is viewed as the testing set and the remaining participants are used for training and validation. As for the unsynchronized dataset, we could not find the ground truth to evaluate performance since we did not have the time labels for each frame. However, this is closer to practical situations where video and balance board frames are not timestamped. In order to compare the dataset improvement before and after the modification based on the estimated results from the Alignment Network, we proposed a differential network to evaluate the COP moving amplitude between two video sequences from the same participant.

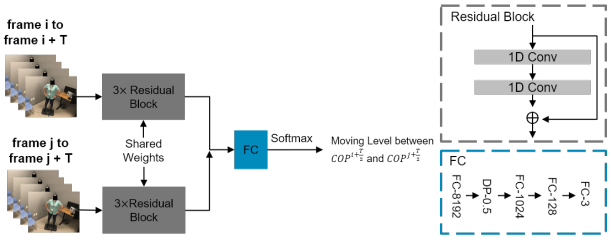


Fig. 6: Structure of the Differential Network [17]. The left figure is the differential network structure. Two video clips (length T = 65) are the input. The output is the moving level determined by the distance between COP locations under two video clips. The right two figures show the structure of the residual block and fully connected layers.

D. Differential Network

Due to each participant having a different standing position on the balance board, the COP origin is not easy to define. We proposed a Siamese differential network [17] that predicted the COP moving amplitude between two input video sequences. The network structure is shown in Fig. 6. The video sequences $S^{i:i+T}_{video}$ (from frame i to frame $i+T$) and $S^{j:j+T}_{video}$ (from frame j to frame $j+T$) were fed into the differential network. The backbones of two branches were combined with three residual blocks. Each residual block was composed of two convolutional layers. The extracted features from two streams were concatenated and used as the input to the fully connected module. The ground truth was acquired based on the COP distance between the corresponding video time stamps $COP^{i+\frac{T}{2}} - COP^{j+\frac{T}{2}}$. Based on the distance value, we were able to apply the corresponding moving amplitude labels (small, moderate and high) to the data. The cross-entropy loss function was used for optimizing the parameters of the Differential Network.

V. EXPERIMENTAL RESULTS

The experimental results are divided into two primary outcomes. The first outcome involved checking the performance of the alignment network on the synchronized dataset by manually set global temporal offset. The second outcome was implemented on the unsynchronized dataset. We aligned the unsynchronized dataset with a trained alignment network from the synchronized one to verify the model's generalization ability. Furthermore, the estimated alignment result was evaluated by a downstream task, a falling risk detection task, with a differential network.

A. Synchronized dataset

Seven participants were included in the synchronized dataset; as previously stated, we used the leave-one-subject-out protocol. Table I shows the aligned results. The AFE of the input was 6.54, with the average global temporal offset set as 2s. After the alignment network, the offset was reduced to 0.312s, i.e., around 84.4% offset was removed.

TABLE I: The performance of the alignment network on the synchronized dataset

Sequence Type	AFE	Time Offset
Input Sequence (Random Offset)	6.54	2.000s
Output Sequence (Reconstruction)	1.02	0.312s
Original Sequence (Ground Truth)	0.00	0.000s

B. Unsynchronized Dataset

Due to the lack of synchronization data, we used a differential network to evaluate the aligned performance on the unsynchronized dataset.

a) Temporal Offset Prediction: We used the trained alignment network from the synchronized dataset to align the unsynchronized dataset. We observe that over 96.4% of the predicted offset value is positive, which matches the fact that the video is taken earlier than the balance board collection.

b) Differential Network Prediction: After correcting the alignment problem, the unsynchronized dataset can be used for some downstream applications. We utilized a differential network that predicted the sway level of COP between two video sequences from the same participant. For comparison, we used the same unsynchronized dataset without predicted alignment information to train the same differential network.

TABLE II: The performance of the alignment network on the unsynchronized dataset. One is without the predicted alignment and the other is with the predicted alignment. The corrected labels after alignment help the network to learn a more accurate model. All three cases achieve improvement in precision and recall metrics with the Alignment Network.

Sequence Type	Precision (Recall)		
	Low	Moderate	High
Sequence without Alignment	0.758 (0.761)	0.703 (0.720)	0.610 (0.580)
Sequence with Alignment	**0.923** (**0.923**)	**0.895** (**0.873**)	**0.805** (**0.847**)
Improvement	21.77% (21.28%)	21.45% (21.25%)	31.97% (46.03%)

The result is shown in Table II. Since the number of Low or Moderate fall risk samples was larger than that of the High-risk samples, they had better performance than the High-risk case. Moreover, with alignment, all three cases gained over 20% improvement due to more accurate ground truth. The varied temporal offsets were largely removed after alignment, which improved the accuracy of the labels used for training. This improvement was especially true for the High-risk case, where the original low-performance result was suffered from both the incorrect labels and limited samples. After alignment, the labels' problem was removed and the gained precision was over 30%.

VI. CONCLUSIONS

In this work, we applied the Alignment Network to solve the synchronization problem between videos and center of pressure sequences. Based on the evaluation results from the two datasets, a significant improvement was obtained after alignment. The removed temporal offset can provide more accurate labels for downstream tasks. However, the requirement of at least one synchronized dataset and adaptation problems between different datasets still restrict the application scenario, which will be studied in future works.

REFERENCES

[1] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103–7112.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[3] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5386–5395.

[4] J. Scott, C. Funk, B. Ravichandran, J. H. Challis, R. T. Collins, and Y. Liu, "From kinematics to dynamics: Estimating center of pressure and base of support from video frames of human motion," *arXiv preprint arXiv:2001.00657*, 2020.

[5] C. Du, S. Graham, S. Jin, C. Depp, and T. Nguyen, "Multi-task center-of-pressure metrics estimation from skeleton using graph convolutional network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2313–2317.

[6] J. Picone, "Fundamentals of speech recognition: A short course," *Institute for Signal and Information Processing, Mississippi State University*, 1996.

[7] F. De la Torre, "A least-squares framework for component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1041–1055, 2012.

[8] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[9] J. Wang, Z. Fang, and H. Zhao, "Alignnet: A unifying approach to audio-visual alignment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3309–3317.

[10] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[11] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4929–4937.

[12] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[13] B. Fischer, V. Roth, and J. M. Buhmann, "Time-series alignment by non-negative multiple generalized canonical correlation analysis," in *BMC bioinformatics*, vol. 8, no. 10. BioMed Central, 2007, pp. 1–10.

[14] J. Lewis, "Automated lip-sync: Background and techniques," *The Journal of Visualization and Computer Animation*, vol. 2, no. 4, pp. 118–122, 1991.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] H.-k. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action-agnostic human pose forecasting," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1423–1432.

[17] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez, "A differential approach for gaze estimation with calibration." in *BMVC*, vol. 2, no. 3, 2018, p. 6.