Interaction Design and Architecture(s) Journal - IxD&A, N.66, 2025, pp. 226 - 245
DOI: 10.55612/s-5002-066-009

Machine Learning-Based Human Action Estimation for Interactive Spatial Lighting Control

İrem Kekilli[1], Muhammet Furkan Yiğitoğlu[1], Orkan Zeynel Güzelci[2]

[1] Istanbul Technical University, Graduate School, Architectural Design Computing Program, Ayazağa Campus, 34496, Istanbul, Türkiye
{kekilli24, yigitoglu24}@itu.edu.tr
[2] Istanbul Technical University, Faculty of Architecture, Department of Interior Architecture, Taşkışla Campus, 34367, Istanbul, Türkiye
guzelci@itu.edu.tr

Abstract. This study investigates the integration of machine learning and architectural lighting design by proposing a proof-of-concept adaptive lighting system driven by human actions and spatial position. A custom video dataset was created based on five actions—standing, sitting, walking, running, and dancing—and three positional categories within a defined space. Two different machine learning approaches were evaluated for human action recognition: a skeleton-based model using MediaPipe pose extraction with an LSTM architecture, and a pixel-based approach combining feature extraction from raw video frames with an MLP classifier. The classified action and position data were mapped to pre-defined lighting schemes generated parametrically using Grasshopper, enabling context-aware lighting recommendations. The results show that while action classification accuracy is limited due to dataset size, position recognition achieves high reliability. The study highlights the potential of action-oriented, human-centered lighting systems and outlines directions for future research involving larger datasets and user-centered evaluations.

Keywords: Machine learning, Human action estimation, Motion-based lighting interaction, Spatial lighting control, Adaptive lighting systems.

1 Introduction.

In architecture, space is not merely a physical environment; it is a structure that gains meaning through the movements and actions of its users and is constantly changing. While the perception of space is shaped by human movement, space also shapes human movement. This relationship is particularly important for human-centered design approaches. Another fundamental element that influences the perception of space is lighting. Therefore, a well-designed lighting system not only provides visual comfort to the user but also determines the person's orientation, emotions, and the atmosphere of the space. For this reason, in the development of both interactive and human-centered architectural design, the perception of movement within space, along with the

development of lighting schemes appropriate to human actions, makes an important contribution.

In this study, movement is not only a physical action but also a fundamental element in the experience of space. The user's position during walking, running, and sitting actions reflects their relationship with the space. In this context, movement not only guides spatial organization but also creates important data for environmental control elements. Therefore, this study lies at the intersection of architecture, interaction design, and intelligent environments, and uses machine learning models (such as LSTM and MLP) to analyze and utilize human movement data. In this way, the system directs an adaptive lighting system using such data. This contribution is consistent with a human-centered and context-aware understanding of spatial control.

In line with this approach, human actions were detected in a defined space, the location of each action was determined, and then a machine learning model was developed to match each action type with specially prepared lighting schemes. The aim is to contribute to the integration of variable and interactive lighting solutions based on user actions into architectural design processes.

A dataset was created by recording five basic actions—standing, sitting, walking, running, and dancing—involving different participants across multiple spatial settings. Trained models were used to identify movements in the space and determine their positional information. Lighting schemes were generated for each movement class and position combination. Subsequently, the action and position information were integrated with the appropriate lighting schemes. The study aims to incorporate artificial intelligence into architectural design processes by primarily focusing on user movements in a space and enabling the creation of interactive and variable lighting scenarios.

The expected outcome of this work is to detect different actions performed in a space and their locations using artificial intelligence and then match them with lighting schemes appropriate to the action types based on these detections. Thus, the role of the user in the design of architectural systems is examined using different algorithms. The contribution of this study to the field is to explore ways to enrich the user experience in different spaces through a motion-oriented lighting approach, conveying the potential of machine learning in the design of environmental control elements in architecture. Therefore, acting as a proof-of-concept, this study demonstrates the feasibility of bridging the gap between human action data and adaptive lighting control systems through machine learning approaches.

This proposed adaptive lighting control system supported by machine learning can be useful from both the designer’s and the user’s perspectives. Even though lighting design is an important component in architecture and interior design practices, it requires specialized expertise that general practitioners may lack. Consequently, architects can utilize this system as a design decision-support tool. It can guide architects in choosing appropriate lighting types, lux values, and color temperatures according to spatial function. From the user’s perspective, in cultural and social spaces (museums, galleries, classrooms, offices, therapy centers, etc.), this adaptive system adjusts lighting parameters to increase user comfort and reduce energy waste. Moreover, this human-centric adaptation enriches personal experiences in spaces dedicated to well-being, such as meditation centers, as well as domestic settings.

2 Literature Review

2.1 Human Motion Capture and Machine Learning

Human motion estimation and recognition have become central research areas in computer vision, robotics, and artificial intelligence. Early approaches typically relied on physics-based models such as constant acceleration or minimum-jerk models, often combined with Kalman filters. However, these methods proved insufficient for providing reliable uncertainty measurements for complex human movements [1].

To overcome these limitations, data-driven models utilizing probabilistic and deep learning techniques have increasingly gained traction. Gaussian Mixture Models (GMM) have been widely used for goal recognition and trajectory estimation. Task-parameterized formulations and online expectation–maximization algorithms have demonstrated GMM's potential for adapting to new trajectories. Hidden Markov Models (HMM) have also emerged as a dominant tool for the stochastic representation of human movements and incremental updates. Furthermore, Dynamic Time Warping (DTW) and generalized time warping methods have been proposed for aligning multimodal sequences, but they have shown limitations in online applications [1].

More recent studies have turned to deep learning models that directly learn spatiotemporal dependencies from skeletal data. For example, Bütepage et al. [2] presented a deep representation learning framework based on an encoder–decoder architecture and demonstrated that it can generalize well even to unseen movements. Martinez et al. [3] examined recurrent neural networks (RNNs) for human motion prediction, revealing that LSTMs and GRUs can capture temporal dependencies, but that simpler approaches sometimes perform better for short-term predictions. Similarly, Kanpak and Arserim [4] applied deep learning methods for human pose estimation, highlighting the importance of joint detection and skeleton-based modeling for accurate classification in complex spatial contexts. Chen and Xue [5] showed that a CNN trained on tri-axial accelerometer time-series data can automatically learn features and achieve approximately 93.8% accuracy, thereby removing the need for manual feature engineering and simplifying the mobile pipeline. In another study in the field of human motion detection, Song et al. [6] recorded lower-limb movements with a portable sEMG system—a method that records muscle electrical activity with electrodes on the skin—extracted simple time-domain and frequency-domain features, and compared MLP and LSTM models. They found the best results (MLP ≈ 95.5%, LSTM ≈ 96.6%) using combined time- and frequency-domain features [6].

Beyond classification, regression and machine learning approaches based on skeleton data have also been used in predictive modeling. A recent study demonstrated that various algorithms, such as Support Vector Machines (SVM), Random Forests, and deep neural networks, are effective in predicting human movements based on skeletal data and are applicable in fields like virtual reality and security systems [7]. Collectively, these studies reveal the evolution from manually extracted features to deep learning frameworks that can better capture the variability and complexity of human movement.

2.2 Lighting and Machine Learning Applications

Parallel to developments in motion prediction, research has also been conducted on integrating human activity recognition into smart lighting control systems. The basic principle is that appropriate lighting conditions, including color temperature, depend on human activities, and therefore these systems need to recognize activities and provide an adaptive response.

Chun and Lee [8] proposed an intelligent lighting control framework based on motion tracking using depth and thermal cameras. This system was able to control lighting levels and color temperature based on real-time activity prediction, aiming to achieve energy savings and user comfort simultaneously. Building on this work, Chun et al. [9] developed a real-time lighting control method that uses multiple camera systems and determines human location from depth images through inverse perspective mapping. The system recognized different activities such as working, chatting, and watching television, providing appropriate lighting conditions while also utilizing the energy efficiency advantages of LED-based lighting.

These approaches emphasize both energy efficiency and user comfort. For example, activity-based LED lighting systems not only reduce energy consumption but also enhance user well-being by adjusting the color temperature to suit different tasks [9]. Distributed lighting systems and gaze direction detection further increase adaptability, enabling context-aware control in living spaces.

Gopalakrishna et al. [2] developed context-based intelligent lighting models for “breakout areas” in office environments. Synthetic data were generated using a probabilistic model that considered six features: user identity, activity type, number of users, activity area, time of day, and external light conditions. Various classification algorithms were then tested, and the study noted that the DecisionTable algorithm achieved the best performance in predicting user preferences.

Putrada et al. [10] reviewed the smart lighting literature and emphasized that machine learning methods play an important role in improving user comfort. They reported that supervised learning, clustering (K-means, DBSCAN), deep learning (CNN), and reinforcement learning methods have been applied in different contexts. Parameters such as light usage rate, unmet comfort rate, Kruithof’s comfort curve, correlated color temperature, and flicker perception were used to measure user comfort.

A review of the literature generally shows that machine learning applications in smart lighting systems offer significant contributions to both energy efficiency and user comfort. Camera- and sensor-based activity recognition approaches [5,8], classification models based on contextual data [2], and supervised, unsupervised, and reinforcement learning methods highlighted in systematic literature reviews [10] reveal diverse orientations in this field. Collectively, these studies indicate that integrated lighting solutions that more accurately predict user behavior while simultaneously optimizing comfort and energy savings can be developed in the future.

2.3 Research Gap

Although significant progress has been made in the fields of human action prediction and smart lighting, research examining the integration of these two areas remains

limited. Studies on motion prediction have mostly focused on trajectory prediction and activity recognition in domains such as robotics, security, and human–computer interaction [1,11]. Similarly, smart lighting research has generally concentrated on inferring broad activity modes using sensor networks or vision-based tracking methods [8,9]. However, studies that directly combine detailed motion classification with dynamic lighting control remain scarce.

This gap highlights the novelty of approaches that explicitly link human motion recognition and location tracking to machine learning–based adaptive lighting systems. By bringing these two fields together, smart environments capable of responding to human behavior in real time while also optimizing energy performance may become feasible in future applications.

3 Methods

A two-stage method was applied to both models used in the study. In this method, a machine learning model was first trained with a dataset grouped by action class. Then, the lighting schemes corresponding to this action class were selected, and the results were generated. As a preliminary test, experiments were conducted with various publicly available datasets. In previous studies, the UCF101 [12] and UTD-MHAD [13] datasets, as well as videos from Pexels [14], were used. Variations in these datasets—such as background differences, frame sizes, and whether the entire action was captured within the frame—prevented the trained model from accurately predicting motion in the video and resulted in significant differences in training times.

These preliminary experiments led to the establishment of dataset criteria. According to these criteria, a fixed background and multiple variations of the same action contributed positively to the model's accuracy and learning efficiency.

For the main experiment, a dataset was prepared based on these. Two different models were trained using this dataset. The aim was to develop deep learning models that simultaneously predict both physical action (pose classification) and spatial position. This section covers the preparation and augmentation of video data, the preparation of lighting schemes, the training of machine learning models, and the prediction process using two different approaches (Fig. 1).

3.1 Data Preparation and Augmentation

Before the model training process, a data preparation stage was carried out. For five different actions (walking, standing, sitting, running, and dancing), 36 videos were recorded for each action, totaling 180 videos. All videos were recorded by the researchers and are original to this study. For each action, videos were captured at left, center, and right positions. The videos feature three participants to introduce variation in the learning process. Left- and right-positioned videos were mirrored for data augmentation, increasing the total number to 270 videos (Fig. 2).

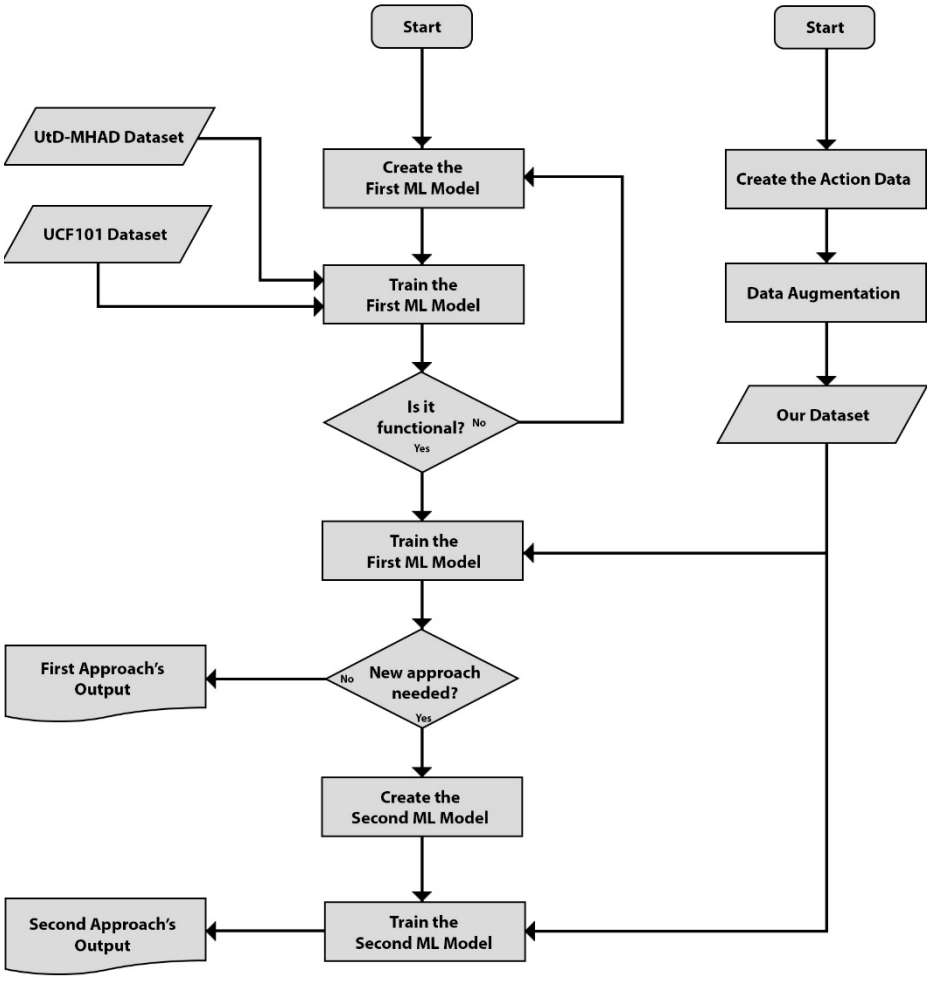

Fig.1. Workflow of the proposed method.

After these steps, the dataset was defined as a 15-class classification problem consisting of five different actions (dancing, running, sitting, standing, walking) and three different position categories (center, right, left). These classes were labeled manually. Interactive machine learning is normally framed in terms of supervised learning, a subclass of machine learning problems in which the computer is presented with a number of examples, each of which has a "label" representing the correct output of the system [15]. For example, labels such as "Running_Left" or "Sitting_Center" were used. Under each class directory, there were videos with .mp4 or .avi extensions (each video was approximately 5 seconds long). These class names were defined at the beginning of the code and set as a total of 15 tags.

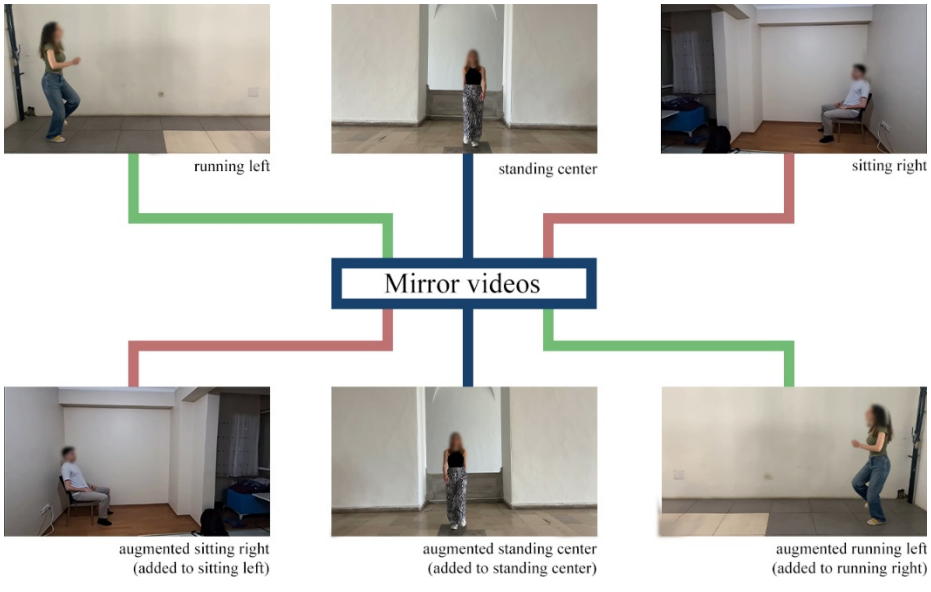

Fig.2. Data augmentation workflow and class labeling.

3.2 Lighting Proposal

The machine learning part concludes with the classification of the action and position in a video. Then, pre-prepared lighting schemes are automatically selected based on the classified action class and positional information. For each action–location combination, a specific lighting view is prepared (Fig. 3).

The light intensity and illuminated area vary for each motion class. Depending on whether the detected motion occurs in the center, right, or left of the space, a cross-sectional view of the corresponding lighting scheme is presented to the user. In other words, the system ultimately provides the user with the following output: “In this video, {movement class} is being performed, and the person is in {position}. Therefore, the recommended lighting scheme is as follows.” The system then displays the relevant lighting diagram.

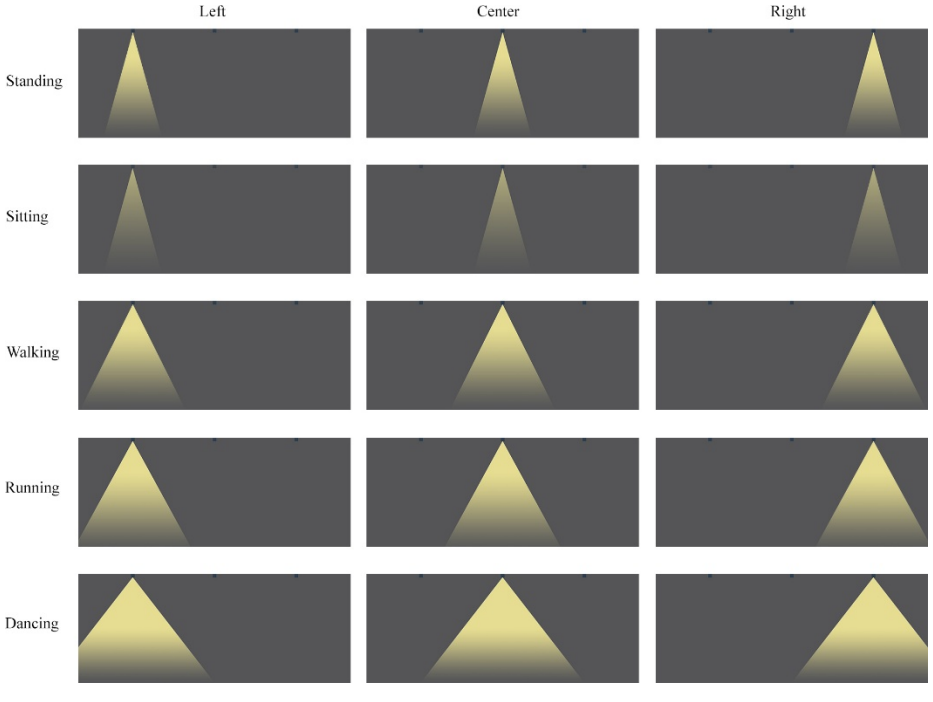

Fig. 3. Lighting schemes prepared for each action–location combination.

3.3 Motion Classification Using Skeleton Data (First Approach)

The first approach develops a deep learning model that simultaneously estimates both the physical action (pose classification) and the position of an individual on the screen using skeleton-based coordinates extracted from video data. Skeleton data were extracted using MediaPipe, an open-source framework developed by Google. This pose estimation solution focuses on real-time, high-fidelity body tracking and is designed to enable data inference from sensory inputs such as video streams or photos, making it suitable for rapidly prototyping perception pipelines. Due to its versatility, MediaPipe Pose is accessible for use within web environments, mobile applications, and across various platforms [16].

The model is built on a multi-task, multi-layered LSTM-based architecture. LSTM (Long Short-Term Memory) is a specialized type of recurrent neural network (RNN) designed to work with sequential data. It can process past information without forgetting it and provides robust results for time-dependent inputs. By effectively learning temporal dependencies among video frames, this architecture captures the geometric relationships of joints over time and increases the stability of joint predictions for moving bodies [17]. Compared to standard RNNs, LSTM architectures are particularly effective in modeling sequential data.

Data Processing for the First Approach. Thirty frames were extracted from each video, resulting in a total of 2,700 × 30 = 81,000 frames. The entire process of framing, extraction, labeling, training, and estimation was carried out on Google Colab using Python code.

Action and Position Estimation for the First Approach. Videos were collected in five different movement classes: walking, standing, sitting, running, and dancing. Three position categories were created for each class: left, center, and right. Each video was represented by 30 fixed frames. Thirty-three body points were detected in each frame using the MediaPipe Pose library (Fig. 4; Fig. 5). The videos were then labeled for action (0 = walking, 1 = standing, 2 = sitting, 3 = running, 4 = dancing) and position (0 = left, 1 = center, 2 = right) categories. The labels were assigned according to file names and checked manually.

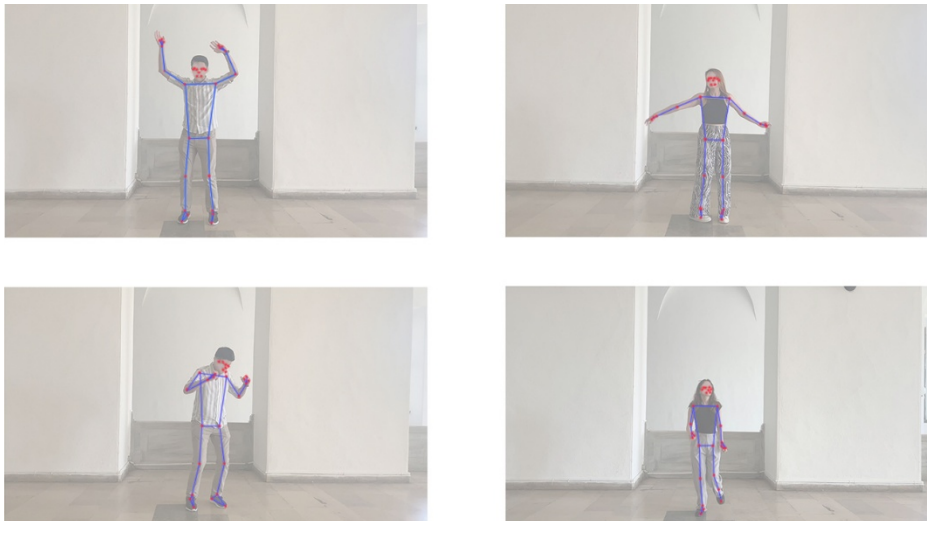

Fig. 4. Skeleton joint detection using the MediaPipe Pose library.

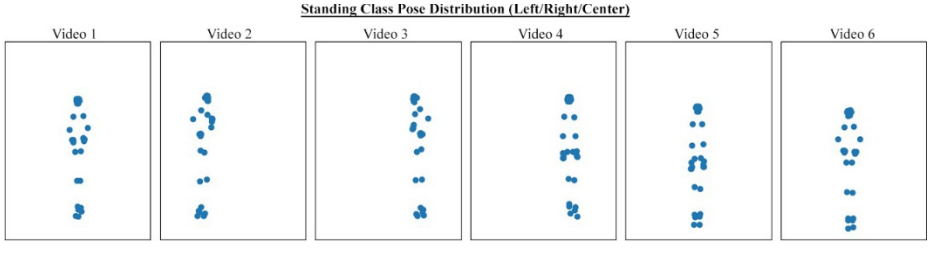

Fig. 5. Skeleton joints detected during the standing pose.

The model structure consists of two parts: a pose estimation module and a position estimation module. Since body movements unfold as sequential data, a multi-layered LSTM architecture was employed for accurate temporal modeling.

The bidirectional LSTM layer allows sequences to be read from both directions. Dropout layers with a rate of 30% were added to prevent overfitting. ReLU was used as the activation function in the hidden layers, while Softmax was employed for multi-class prediction in the output layers. During training, the sparse categorical cross-entropy loss function and the Adam optimization algorithm were used. The architecture of the proposed model is summarized in the code snippet shown below.

```
x=Bidirectional(LSTM(64,return_sequences=True))(inputs)
x=Dropout(0.3)(x)
x=LSTM(64,return_sequences=True)(x)
x=Dropout(0.3)(x)
x=LSTM(32)(x)
x=Dropout(0.3)(x)
x=Dense(64,activation='relu')(x)
pose_output=Dense(5,activation='softmax',name='pose_out
put')(x)
position_output=Dense(3,activation='softmax',name='posi
tion_output')(x)
model.compile(
  optimizer=Adam(learning_rate=0.001),
  loss={
    'pose_output':'sparse_categorical_crossentropy',
    'position_output':'sparse_categorical_crossentropy'
  },
  metrics={
    'pose_output':'accuracy',
    'position_output':'accuracy'
  }
```

The dataset was divided into 80% for training and 20% for validation. During training, class weights were calculated and incorporated into the model to prevent class imbalance. Training was optimized with a batch size of 16 and for 100 epochs.

With this structure, both human actions and their positional information in the x-plane were estimated simultaneously. The model provides balanced information sharing between the two tasks, particularly due to the separate output layers following the ReLU layer. Since the motion and position models operate independently, each produces separate predictions.

3.3 Motion Classification Using Raw Video Images (Second Approach)

In addition to motion extraction using human skeletons, motion classification was also performed using raw video images. One of the main challenges in this approach was the high computational and memory requirements. This section describes the methods and optimization steps used to address these limitations.

Using CNN + RNN for the Second Approach. Convolutional Neural Networks (CNNs) are deep learning architectures commonly applied in image and video recognition tasks. A CNN typically consists of three types of layers: convolutional, subsampling (pooling), and fully connected layers [18]. The architecture of a CNN is organized as a series of stages [19]. The convolutional and pooling layers apply local receptive fields and shared weights, and they can be stacked into multiple layers. Classification is performed through a fully connected layer at the final stage [20]. CNNs have advantages in local feature learning but disadvantages in temporal modeling and in learning effectively from small datasets. They are widely used in domains such as image classification (ImageNet), object detection (YOLO), segmentation (U-Net), and style transfer.

Recurrent Neural Networks (RNNs) are a family of neural networks designed to capture sequential dependencies in time series, text, or other sequential data. They learn the features of temporal sequences through the memory of previous inputs in the internal state of the network [20]. At each step, the hidden state combines the previous memory with the new input, making decisions based on past information. Although they can process sequences of variable length, their step-by-step dependency leads to high GPU utilization. RNNs are commonly used in natural language processing, speech recognition and synthesis, time series prediction, and machine translation [18].

To extract features from each video frame, a pre-trained CNN similar to ResNet50 was employed. From each frame, a 2048-dimensional feature vector was extracted. These features were then passed to a sequential LSTM layer to model motion along the temporal axis, aiming to capture the evolution of movement. However, this method imposed heavy computational and memory demands, as the entire sequence of frames was processed in every training iteration.

To reduce computational cost, the input images were scaled and converted to a smaller resolution (112 × 112). In some experiments, the initial layers of the model were frozen (i.e., not retrained) to further decrease the computational load. Additionally, training data were reduced by selecting only frames from key moments of each video. Although these optimizations shortened training times, they also caused a loss of accuracy in predicting action classes and positions due to the limited size of the dataset.

Feature Extraction and MLP for the Second Approach. A multilayer perceptron (MLP) network, which is based on a feedforward artificial neural network, was used for supervised learning and classification [21]. An MLP consists of an input layer for the initial data, one or more hidden layers to capture complex patterns, and an output layer to generate the final prediction [22].

In the first step, Keras' pre-trained ResNet50 model was employed to extract features (feature vectors) from the videos. By removing the post-classification layers of the model, 2048-dimensional feature vectors were extracted from the final pre-classification layer. Each time a video frame was fed into ResNet50, a feature vector was generated.

Step-by-step, ResNet50 pre-trained with ImageNet data was loaded; its output was a 2048-dimensional vector for each frame. The model functioned solely as a feature extractor. For each video, the video_to_vector function was applied, with OpenCV reading the frames one by one. Each frame was rescaled to (IMG_SIZE, IMG_SIZE)

(e.g., 112 × 112), and pixel values were normalized to the range 0–1. The selected frames were passed through the feature extractor layer of the model, resulting in a 2048-dimensional vector for each frame. These vectors were averaged to produce a single 2048-dimensional fixed-size vector per video. The resulting feature vectors and the corresponding class labels were saved to disk in compressed .npz format, with one file generated for each video. In this way, precomputed features for the entire dataset were stored.

This approach allowed the features to be precomputed and stored in advance. Since the extracted features were saved to disk, they did not need to be recomputed during each training iteration, which significantly reduced the computational cost.

MLP Model and Training for the Second Approach. Pre-extracted features were stored to create the input array x and the label array y. These arrays were converted to NumPy arrays (float32 for X, int32 for y). The data were split into 80% for training and 20% for validation.

A simple Multilayer Perceptron (MLP) model was defined for training. The structure of the model was as follows:

- Input layer. The dimension was feature_dim (e.g., 2048), representing the feature vector extracted for each video.
- Intermediate layer 1. A dense layer with 256 neurons (ReLU activation), followed by a 30% dropout layer to prevent overfitting.
- Intermediate layer 2. A dense layer with 128 neurons (ReLU activation), followed by a 30% dropout layer.
- Output layer. A dense layer with 15 neurons (Softmax activation), providing class probabilities.

4 Results

4.1 Findings for the First Approach

As a result of training, it was observed that position classification and estimation were successful. The action training graph shows that the model was in the process of learning, but the learning was insufficient (accuracy < 0.95). Due to the limited number of videos, the learning curve fluctuated between 0.4 and 0.6 (Fig. 6).

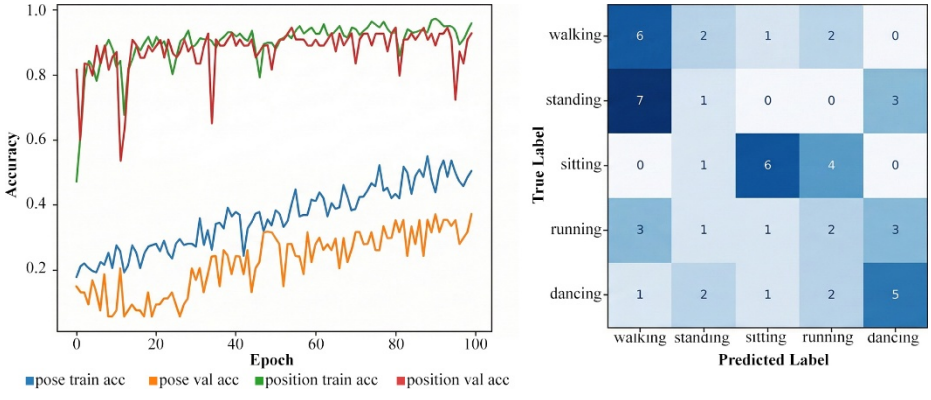

Fig. 6. (Left) Training and validation curves for pose and position estimation; (Right) confusion matrix for pose estimation.

In the action prediction confusion matrix, the videos were mostly classified correctly in the Walking, Sitting, and Dancing classes, although some misclassifications occurred (Fig. 6). The Standing and Walking classes were often confused due to similar body postures, and a similar situation was observed between the Running and Dancing classes (Fig. 7; Fig. 8). In the position prediction confusion matrix, the action positions were predicted correctly in most cases, with only four misclassifications out of 54 videos.

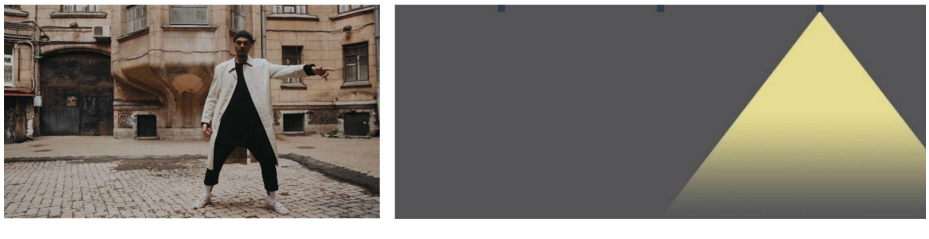

Fig. 7. (Left) Ground-truth video: Dancing, Right [14]; (Right) model prediction and corresponding lighting proposal: Dancing, Right.

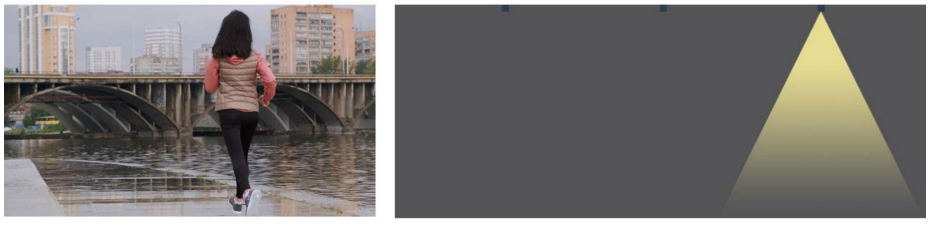

Fig. 8. (Left) Ground-truth video: Running, Right [14]; (Right) model prediction and corresponding lighting proposal: Walking, Right.

4.2 Findings for the Second Approach

In the early experiments, training took a very long time (e.g., ~1 hour per epoch) because the model repeated all CNN operations in each epoch. In the current method, since the features were pre-extracted, only small-sized feature vectors were trained in each epoch, which reduced the computational burden. As mentioned earlier, precomputing the output of the pre-trained layers and saving them to disk eliminated the need to repeat this process during every training round. As a result, an epoch that previously took ~1 hour was reduced to only 2–3 seconds. This speedup was a direct result of completing the feature extraction step prior to training. In this way, higher epoch numbers were achieved, and more reliable results were obtained. After training for 500 epochs, the model reached an accuracy of 0.7132 with a loss of 0.7244, while the validation accuracy was 0.6032 and the validation loss was 1.0403 (Fig. 9).

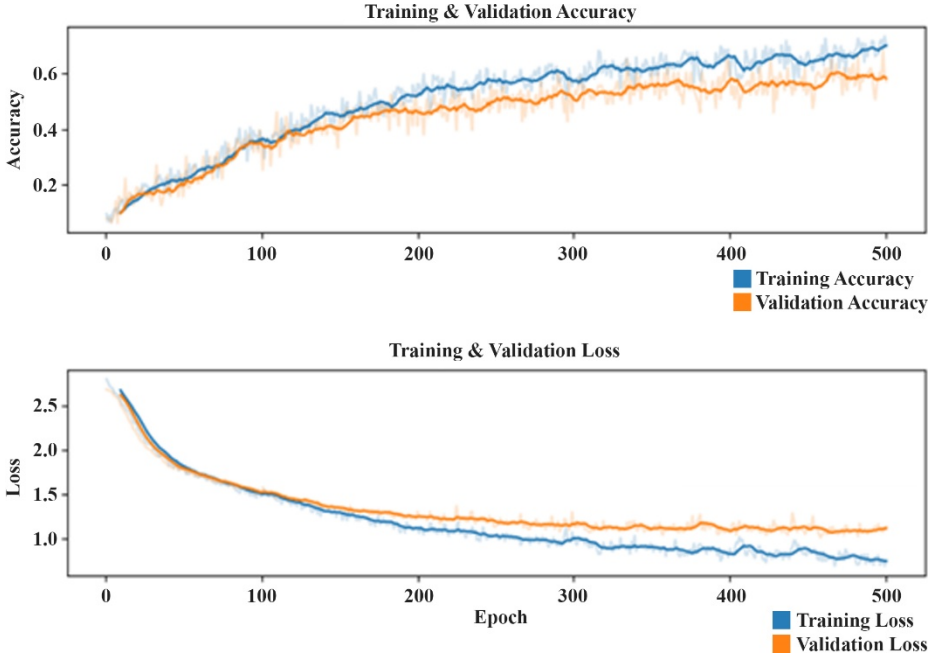

Fig. 9. Training and validation metrics for the feature extraction + MLP approach.

4.3 Comparison of Results from Two Approaches

In general, the first model's position classification achieved high accuracy in both the training and validation sets (Table 1). In contrast, pose classification showed signs of overfitting, with a notable gap between training and validation accuracy.

Table 1. Loss and accuracy results for the training and validation sets (first approach).

Metric	Training Set	Validation Set
Pose Accuracy	72.5%	38.8%
Position Accuracy	97.2%	92.6%
Pose Loss (Final Epoch)	0.8146	1.7386
Position Loss (Final Epoch)	0.1123	0.1293
Total Loss (Final Epoch)	0.9269	1.8679

The pose confusion matrix indicated misclassifications between the Walking and Standing classes, whereas the Sitting and Dancing classes were predicted with higher reliability due to their more distinguishable actions.

In the second approach (motion classification using raw video images), since the use of CNN + LSTM significantly extended the epoch durations, the features of the dataset were extracted in compressed .npz format prior to training, and the training was performed on these precomputed features. As a result, epoch durations were reduced from approximately one hour to only 2–3 seconds, greatly accelerating the training process.

Since the model in the second approach was trained for a single combined (motion–position) classification task, the separate "Pose Accuracy," "Position Accuracy," "Pose Loss," and "Position Loss" metrics that were present in the first approach are not available. The values shown therefore correspond to the combined classification of pose and position (Table 2).

Table 2. Loss and accuracy results for the training and validation sets (second approach).

Metric	Training Set	Validation Set
Pose-Position Accuracy	75.3%	63.5%
Pose-Position Loss	0.648	0.986

Video predictions achieved confidence values of up to 0.90. Confidence decreased for actions with similar characteristics, such as Walking and Running, whereas actions such as Sitting achieved particularly high confidence scores. For similar motions (e.g., Walking vs. Running), some instances were misclassified in terms of motion class, but the location prediction remained correct (Fig. 10).

Fig. 10. Predictions obtained using the feature extraction and MLP approach.

The "Sitting" class achieved the highest performance with an F1-score of 0.83 and a notably high recall of 0.92. This is a critical finding for the lighting system, as "Sitting" represents a state in which users are likely to remain stationary for extended periods (e.g., reading, resting), requiring stable and consistent illumination. The model successfully identified this state in 92% of cases.

The lowest performance was observed in the "Walking" (F1: 0.36) and "Standing" (F1: 0.43) action classes. This confirms the visual ambiguity mentioned earlier; the transition frames between walking and standing are morphologically similar, causing the model to confuse these labels.

While the overall pose accuracy was 0.52, the high success rate in "Sitting" ensures that the system performs well in the most duration-heavy activity. For more dynamic and ambiguous classes, such as "Walking" vs. "Standing," the system's fail-safe relies on the position accuracy (97.2%), ensuring that even if the action is misclassified, the lighting location remains correct (Table 3).

The fact that positional accuracy remained high in all tests demonstrates that the system reliably determines the position of a person in the space, even if it occasionally misinterprets the nuances of dynamic movement.

Table 3. Precision, recall, and F1-score results for the second approach.

Action Class	Precision	Recall	F1-Score
Sitting	0.75	0.92	0.83
Dancing	0.39	0.58	0.47
Running	0.56	0.38	0.45
Standing	0.50	0.38	0.43
Walking	0.40	0.33	0.36

5 Discussion and Conclusion

This study aims to develop a spatial lighting experience through body actions. In interior lighting design, factors such as visual performance and visual comfort are central, as delineated in the lighting standards set by the International Commission on Illumination [23]. Instead of a fixed spatial lighting system, a human action–adaptive lighting control system enhances the spatial experience through interactivity. Within this approach, there are two main objectives: (i) to obtain a lighting output for a space according to the experiencer's action and position by using machine learning models, and (ii) to compare skeleton-based and pixel-based machine learning models to identify the best results for human action recognition.

This study demonstrated that it is possible to use body action as an input for spatial lighting. Even though the results of action estimation were insufficient, position recognition produced highly accurate outputs. The findings suggest that further development is required to effectively use body interactions for lighting design. Within the scope of this experiment, position estimation results met expectations, whereas action estimation was insufficient to fully realize the proposed interaction framework.

The limited sample size of the dataset (only 270 videos) made training deep CNN+LSTM models challenging in terms of both time and computational resources. As a proof-of-concept addressing data limitations, a custom dataset was created to avoid background clutter commonly found in publicly available datasets, and no open-source datasets were used. The primary aim of this decision was to build and validate a methodological framework rather than to achieve maximal classification accuracy. Accordingly, this study is positioned as a comparative methodological investigation of two machine learning approaches. Therefore, even though a higher number of classes could improve model performance, five action classes were sufficient to demonstrate the comparative results and applicability of the proposed framework. With the feature extraction + MLP approach, which used high-level vectors extracted from the pre-trained ResNet50, the training time per epoch was reduced from approximately one hour to only a few seconds, while achieving satisfactory accuracy. In comparison, the first approach achieved fast epoch times because it directly processed low-dimensional data from MediaPipe-based skeleton extraction, whereas the second approach provided stronger generalization even with a limited dataset by holistically capturing action and position through the rich information contained in raw images.

In this study, the methods employed reflect a combination of techniques that were available and developed during the research period (February-June 2025). Future research will proceed according to the methods and technologies available at that time. Increasing the dataset size will be essential for improving accuracy, as data limitations in the present study directly affected model performance. Adjusting the architecture and the number of layers of the machine learning models may also yield better outcomes. Furthermore, while the outputs of this study were two-dimensional, future research could explore three-dimensional action recognition, enabling the generation of 3D spatial lighting proposals for interior environments. Moreover, expanding the number of recognized actions would enhance usability across a wider range of activities and spatial contexts, allowing lighting responses to be tailored to diverse environments.

From an architectural design perspective, this machine learning–supported method assists designers in lighting selection based on spatial and user requirements, whether implemented as a fixed lighting scheme or as an adaptive system. By recognizing human actions and modulating the atmosphere in real time, the system transforms space from a static condition into a dynamic experiential environment. Consequently, lighting transcends its conventional role as a utility or decorative element and becomes a flexible design instrument that adapts spatial function according to user activity. This approach enables designers to integrate user experience (UX) principles into physical space, supporting the creation of more personalized, intuitive, and human-centered interiors.

From a methodological standpoint, this study contributes to the intersection of intelligent environments, interaction design, and architecture by establishing an interdisciplinary bridge between computer vision and lighting design. Rather than focusing solely on object recognition, the proposed workflow incorporates the semantic analysis of complex human movements and translates this data into spatial outputs in the form of lighting scenarios. In this respect, the study offers a contribution to the literature on context-aware systems, particularly within the built environment.

In terms of societal implications, action-oriented lighting systems support a vision of sustainable environments by optimizing energy consumption and reducing unnecessary energy use. By adapting lighting conditions to actual occupancy and activity patterns, such systems can contribute to both environmental responsibility and user comfort.

Overall, the findings provide a foundation for future research on body action recognition and interactive spatial systems. Beyond lighting applications, body action estimation holds potential for controlling sound, heating, and other physical or digital environmental parameters. Such extensions may expand the representational and experiential dimensions of architecture. Understanding human behavior and bodily interaction with space deepens spatial perception, and increased interactivity between the body and the built environment can significantly enhance overall spatial experience.

Acknowledgments. This study was developed by the first and second authors within the scope of the Machine Learning for Architectural Design course conducted by the third author in the Architectural Design Computing graduate program at Istanbul Technical University. The authors would like to thank Büşra Şen for her contributions during the early stages of the study.

CRediT author statement.
İrem Kekilli: Conceptualization, Methodology, Data Curation, Validation, Visualization, Writing – original draft. **Muhammet Furkan Yiğitoğlu:** Conceptualization, Methodology, Data Curation, Validation, Visualization, Writing – original draft. **Orkan Zeynel Güzelci:** Conceptualization, Supervision, Writing – review and editing.

References

1. Li, Q., Zhang, Z., You, Y., Mu, Y., & Feng, C. (2020). Data driven models for human motion prediction in human-robot collaboration. IEEE Access, 8, 227690-227702. https://doi.org/10.1109/ACCESS.2020.3045994
2. Gopalakrishna A. K., Özçelebi T., Liotta A., Lukkien J. J.: Exploiting machine learning for intelligent room lighting applications. In: Proceedings of the 6th IEEE International Conference on Intelligent Systems (IS 2012), pp. 406--411. IEEE (2012). https://doi.org/10.1109/IS.2012.6335169
3. Martinez J., Black M. J., Romero J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891--2900 (2017). https://doi.org/10.1109/CVPR.2017.497
4. Kanpak H. N., Arserim M. A.: Human posture prediction by deep learning. Dicle University Journal of Engineering, 12(5), pp. 775--782 (2021). https://doi.org/10.24012/dumf.1051429
5. Chen Y., Xue Y.: A deep learning approach to human activity recognition based on single accelerometer. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1488--1493. IEEE (2015). https://doi.org/10.1109/SMC.2015.263
6. Song J., Zhu A., Tu Y., Huang H., Arif M. A., Shen Z., Zhang X., Cao G.: Effects of different feature parameters of sEMG on human motion pattern recognition using multilayer perceptrons and LSTM neural networks. Applied Sciences, 10, 3358 (2020). https://doi.org/10.3390/app10103358
7. Safibullayevna, B. S., Khanatovna, K. D., Karamatdinkizi, J. M., Faxriddinovich, S. F., & Khairullayevna, S. A. (2024, May). Regression and Machine Learning Methods for Predicting Human Movements Based on Skeletal Data. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-7). IEEE. https://doi.org/10.1109/SIST61555.2024.10629231
8. Chun S. Y., Lee C. S.: Applications of human motion tracking: Smart lighting control. In: IEEE CVPR Workshops, pp. 387-392 (2013). https://doi.org/10.1109/CVPRW.2013.65
9. Chun S. Y., Lee C. S., Jang J. S.: Real-time smart lighting control using human motion tracking from depth camera. Journal of Real-Time Image Processing, 10(4), pp. 805--820 (2015). https://doi.org/10.1007/s11554-014-0414-1
10. Putrada A. G., Abdurohman M., Perdana D., Nuha H.: Machine learning methods in smart lighting toward achieving user comfort: A survey. IEEE Access, 10, pp. 45137--45176 (2022). https://doi.org/10.1109/ACCESS.2022.3169765
11. Bütepage J., Black M. J., Kragic D., Kjellström H.: Deep representation learning for human motion prediction and classification. Computer Vision and Image Understanding, 144, pp. 14--26 (2017). https://doi.org/10.1109/CVPR.2017.173
12. UCF101 – Action Recognition Data Set. (n.d.). Center for Research in Computer Vision, University of Central Florida. https://www.crcv.ucf.edu/data/UCF101.php
13. UTD Multimodal Human Action Dataset (UTD-MHAD). (2015). University of Texas at Dallas. https://personal.utdallas.edu/~kehtar/UTD-MHAD.html

14. Pexels: Pexels. Stock video of human motion used for educational purposes [Video]. https://www.pexels.com/
15. Gillies M.: Understanding the role of interactive machine learning in movement interaction design. ACM Transactions on Computer-Human Interaction, 26(1), Article 5, 1--34 (2019). https://doi.org/10.1145/3287307
16. Roggio F., Trovato B., Sortino M., Musumeci G.: A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review. Heliyon, 10(21), e39977 (2024). https://doi.org/10.1016/j.heliyon.2024.e39977
17. Luo Y., Ren J., Wang Z., Sun W., Pan J., Liu J., Pang J., Lin L.: LSTM pose machines. arXiv preprint arXiv:1712.06316 (2018). https://doi.org/10.48550/arXiv.1712.06316
18. Zhang Q., Yang L. T., Chen Z., Li P.: A survey on deep learning for big data. Information Fusion, 42, pp. 146--157 (2018). https://doi.org/10.1016/j.inffus.2017.10.006
19. LeCun Y., Bengio Y., Hinton G.: Deep learning. Nature, 521(7553), pp. 436--444 (2015). https://doi.org/10.1038/nature14539
20. Yu J., de Antonio A., Villalba-Mora E.: Deep learning (CNN, RNN) applications for smart homes: A systematic review. Computers, 11(2), 26 (2022). https://doi.org/10.3390/computers11020026
21. Talukdar J., Mehta B.: Human action recognition system using good features and multilayer perceptron network. In: Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), pp. 317--323. IEEE (2017). https://arxiv.org/abs/1708.06794
22. Naseer A., Jalal A.: Multimodal objects categorization by fusing GMM and multi-layer perceptron. In: 2024 5th International Conference on Advancements in Computational Sciences (ICACS), pp. 97--103. IEEE (2024). https://doi.org/10.1109/ICACS60934.2024.10473242
23. Liu J., Lou J., Zheng Y., Zhou K.: Automatic indoor lighting generation driven by human activity learned from virtual experience. In: Proceedings of the 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 276--285. IEEE (2024). https://doi.org/10.1109/VR58804.2024.00050