ELEMENTAL MOTION IN SPATIAL INTERACTION (EMSI)

A Framework for Understanding Space through Movement and Computer Vision

MARGARET Z. ZHOU[1], SHI YU CHEN[2] and JOSE L. GARCÍA DEL CASTILLO Y LÓPEZ[3]
[1,2,3] Harvard University.
1margaretzhzhou@gmail.com, 0000-0003-1187-9025
[2]jessica.shiyu@gmail.com, 0000-0002-2826-3318
[3]personal@garciadelcastillo.es, 0000-0001-6117-1602

Abstract. Spatial analysis and evaluation are becoming increasingly common as new technologies enable users, designers, and researchers to study spatial motion patterns without relying on manual notations for observations. While ideas related to motion and space have been studied in other fields such as industrial engineering, choreography, and computer science, the understanding of efficiency and quality in architectural spaces through motion has not been widely explored. This research applies techniques in computer vision to analyse human body motion in architectural spaces as a measure of experience and engagement. A taxonomy framework is proposed to categorize human motion components relevant to spatial interactions, for analysis through computer vision. A technical case study developed upon a machine-learning-aided model is used to test a selection of the proposed framework within domestic kitchen environments. This contribution adds further perspective to wider research explorations in the quality, inclusivity, engagement, and efficiency of architectural spaces through computer-aided tools.

Keywords. Pose Estimation; Spatial Evaluation; Architectural Usability; Motion Studies; Computer Vision; SDG 3; SDG 9.

1. Introduction

To experience architecture, one cannot sufficiently understand its spatial ambitions through viewing - space needs to be navigated, explored, and oriented to obtain an informed experience. Spatial perception and evaluation are considered highly subjective and difficult to measure, and existing qualitative studies rely on surveys and interviews to gather data. Feedback is communicated through verbal and emotional experiences that summarize the user's journey through space. However, how a user feels after occupying space is driven by numerous external and internal factors that vary frequently.

POST-CARBON, Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2022, Volume 1, 505-514. © 2022 and published by the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

Building interaction manifests in the form of human movements that occupy space. Each motion is the output of a subconscious, intuitive decision that the human brain makes in a given moment. Although subjective opinions cannot be measured, actions and trends could be identified and used to extrapolate future decisions. The study of human behaviour and cognition within Architecture is an emerging field. Within this, human circulatory and spatial patterns are particularly important for understanding how space is seen, understood, and used.

Usability studies in fields such as human-computer interaction and product design have enabled designers to evaluate their solutions before being used widely; however, in architecture, this is only possible with simulations and studies from built precedents. Understanding motion trends computationally could help expand research strategies for data simulations, how space shapes movement patterns, and the psychology behind wellbeing, emotion, and comfort in the built environment. Identifying outliers and repetitive or arduous motions, paired with additional cognitive studies, could help pinpoint areas of circulation pressure, energy waste, and ease of use.

The role of computation in addressing environmental pressures of the decade is crucial. Research continues to advance academic understanding of spatial and environmental phenomena in a carbon-neutral way, specifically through computational simulations that can maximize efficiency and usability before build. This research positions itself most predominantly with Goal 3 and 9 of the United Nations Sustainable Development Goals. It aims to further understandings of how the built environment affects users to ensure healthy lives and promote wellbeing for all through both physical ease of use and mental health awareness. This further contributes to building resilient infrastructure that can learn from human responses and interaction to increase inclusivity for all user and body types, and promotes innovation through intelligent, adaptable buildings that put the user first throughout all stages of design, before it is built.

This research explores how the analysis of body motion in architectural spaces could be enhanced using computer vision techniques as a potential tool to depict human experience. A taxonomy framework of motion elements is proposed, designed to visually measure an individual's interactions within built environments. Building upon existing research in motion detection, this codification was developed primarily through iterations of observations, and literature analyses. We propose four informational groups that represent facets of motion relative to architectural space: State, Type, Goal, and Texture. A technical case study developed using the machine learning model PoseNet is used to test the viability of a selected portion of the framework within domestic kitchen environments. (Papandreou et al., 2018). Results from the research could help to access further opportunities in studying usability through motion within fields of universal design, usability, and health and wellbeing.

1.1. BACKGROUND

Earliest forms of academic motion study research emerged with the invention of photographic still capture by early cameras that allowed for tracking from one still frame to the other. Studies on motion tracking became popular in the early 20th century, capturing movements that were previously unseen to the human eye through chronophotographic work with animal locomotion such as horses running and human movement (Muybridge, 1901), or by isolating movement from its context and depicting tracking graphically in order to find new patterns (Marey, 1895).

Point tracking methods, while helpful, have limitations: simple 2D data points are digestible for humans identifying patterns yet indecipherable to the machine. Classifications help to group types of movements together for identification and data usage. Frank and Lillian Gilbreth introduced the idea of motion efficiency, where “motion waste” can be saved by reducing ineffective motions (1919). This led to the creation of Therbligs, a taxonomy of primary movements comprised within common handheld tasks, used to identify costly motion within manufacturing and assembly. Gilbreth’s students furthered these studies into the domestic home by measuring space used by families when performing daily tasks with equipment and furniture. (Callaghan and Palmer, 1944).

Movement classifications also exist in fields such as modern dance. Choreographer Rudolf von Laban created the Laban Movement Analysis, which captures the differences in qualities of movement that would be missed in purely formal analyses (Newlove, 1993; Prinsloo et al., 2019). It has been used in applications such as work efficiency and physical rehabilitation (Ajili et al., 2017).

Recent developments in motion capture and tracking bring technologies that greatly advance accessibility, cost, and accuracy. Optical motion capture systems utilize networks of cameras and sensor markers to process high-resolution movement for use in entertainment, such as OptiTrack, and ergonomic studies (Nagymáté and Kiss, 2018; Bortolini et al., 2018). Motion tracking through devices such as HalO Indoor Positioning bring new ideas to denoting circulation in indoor environments; however, studies are limited to the number of wearable devices available to participants and focus on the complete journeys of individual users (Hu and Park, 2017).

Within computer vision research, Bobick (1997) and Moeslund et al. (2001) provide differing frameworks and definitions for identifying actions such as hitting a baseball or playing football through image-based representation using static and dynamic recognition (Mohamed, 2015). More recent taxonomies, as reviewed by Aggarwal and Ryoo (2011) take different approaches to classifying motion and build upon the idea of identifying categories of “gestures” that are broadly applicable to many different types of actions. There are also public datasets available for machine learning models, such as the KTH and Weizmann datasets, which provide videos for general ‘actions’ such as walking or running (Schuldt et al., 2004; Blank et al., 2005).

In architecture, motion extends beyond efficiency and tracking, embodying a greater purpose of fulfilment, enjoyment, and pleasure. Studies into measuring these qualities have been initiated by the idea of enabling the machine to perceive space as a human designer (Peng, 2017). Hirschberg et al. (2006) discuss the topic of motion within the study of architecture but focus on the opportunity for motion to create form, while Fox and Polancic (2012) explore gestural interaction with intelligent architectural environments. While existing explorations into motion identification address different aspects of the problem space through the lens of industrial engineering, computer science, and design, applications to usability in architectural spaces have not been widely explored.

2. Framework

The following framework combines applications of more recent computational techniques to ideas about quality, engagement, and usage of architectural spaces in order to identify and understand motion.

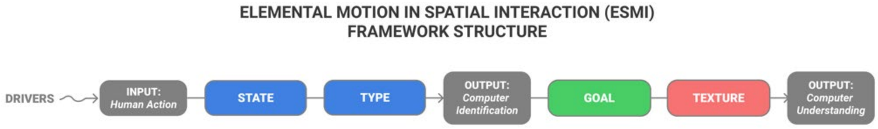
ELEMENTAL MOTION IN SPATIAL INTERACTION (ESMI)
FRAMEWORK STRUCTURE
DRIVERS
INPUT: Human Action
STATE
TYPE
OUTPUT: Computer Identification
GOAL
TEXTURE
OUTPUT: Computer Understanding

Figure 1. Framework Structure - Elemental Motion in Spatial Interaction ESMI

Elemental Motion in Spatial Interaction ESMI is a taxonomy of motion elements for identifying and understanding movements within built environments. Figure 1 describes the structure of the framework in which human action inputs could be identified and then further understood with computer vision. Drawing upon existing research, particularly the framework of Therbligs and those discussed by Aggarwal and Ryoo, it has been developed through observations, discussion, and literature analyses.

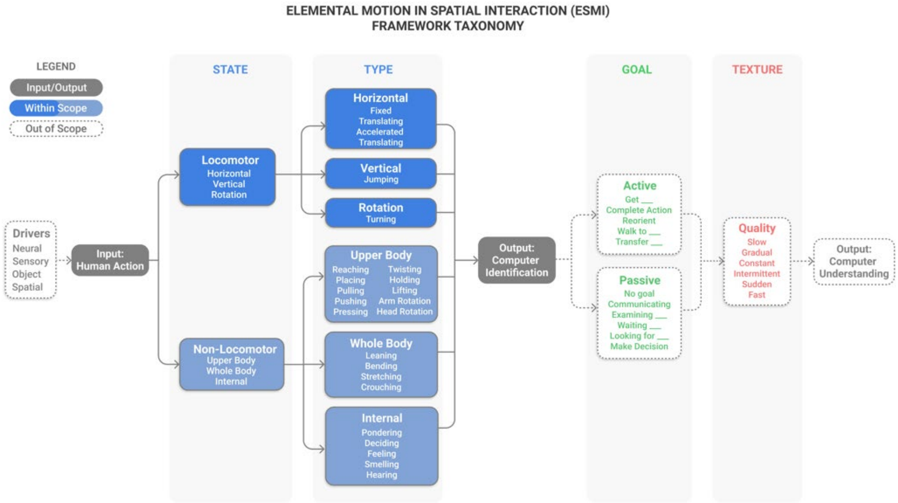

Figure 2. Full Proposed Taxonomy - Elemental Motion in Spatial Interaction ESMI

It proposes four taxonomy groups, *State* and *Type*, which allow an algorithm to firstly identify motion, and *Goal* and *Texture*, which add information for understanding the intent and quality of motion. Figure 2 shows the full framework including lists and hierarchy of identifiable data, and how they build on each other. Both outputs contribute layers of information to the quality of space; the former through data collection of positions of bodies within a given area and respective types of movement, the later through extrapolated data that attempts to understand emotional responses.

Drawing upon Laban's understanding of locomotor and non-locomotor movement, which distinguishes between movements that are done while the body is travelling from point to point versus those that use the axis of the body without travelling, the *State* of a motion is proposed here as one of the first identifications needed when observing motion occurrences (Newlove, 1993). Through determining whether the subject in focus is moving, the taxonomy can provide a general context of what space might be occupied next, as well as what types of movement could be completed. Area used by a person who is standing in the same place, is different from a person who is actively moving across a space.

Type is proposed as a breakdown of primary motion available to a person when using an architectural space. These elements can be layered on top of one another, but each has distinguishable conditions that do not overlap. Locomotor movement is split into horizontal, vertical, and rotational due to their different implications on subsequent, predicted, occupied space. For example, a 90-degree turn has a different prediction to a vertical jump. Non-locomotor is grouped into *Upper Body*, *Whole Body* and *Internal* due to their different implications on purposes. For instance, while a person can be reaching and leaning simultaneously, reaching is mutually exclusive to leaning, and one does not infer the other.

Through processing the *State* and *Type* of a motion, a motion identification could be inferred, noting how the space is being occupied over a period of time. Further contextual groups of information, *Goal* and *Quality*, could help an algorithm understand why it is being used and the impact on users. Both groups infer qualitative factors through physical measurements.

Goal is an identification that gives more insight into the purpose of the motion. Through observations, it was found that an observer would often make predictions to determine what movements might come next or the purpose of the motion initiation. This involves looking at previous trends and predicting future trends under a larger goal. These can be *active* or *passive*. *Active* goals engage with the surrounding spatial context through a physical process, such as walking to another point or transferring an object from point to point. *Passive* goals are typically engaged through an internal process that have physical motions as a by-product, such as tapping the floor while waiting for something. An algorithm can infer this through previous actions, direction, and trend assumptions.

The group of information *Texture* adds qualitative insight into the inference of how an individual might be feeling whilst moving. Observations showed evidence that the speed and style of motion can often relate to emotional states. Laban labels this as effort/dynamics – subtle movements that provide character. For example, a person who is angry is more likely to be tense and contracted. On the architectural scale, this translates into the speed of motion. A person walking at a faster pace is likely determined to achieve something and in a more focused mood. A person moving intermittently may be confused or struggling to find their next goal.

The output of this framework is, firstly, an identification of motion that can quantify how a space is being used over a period, and secondly, an understanding of response to the experience of a space. An example of how the framework could be applied to various scenarios is shown in Figure 3. This contributes to the field of occupancy and ethnographic studies, as well overall usability studies for how individuals respond to specific architectural decisions within spatial environments, potentially revealing measurable trends through the data.

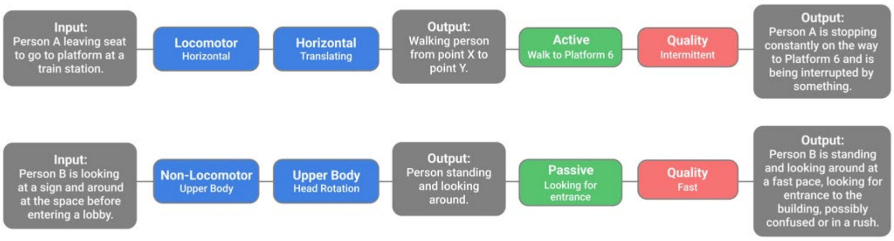
Input:
Person A leaving seat to go to platform at a train station.
Locomotor
Horizontal
Horizontal
Translating
Output:
Walking person from point X to point Y.
Active
Walk to Platform 6
Quality
Intermittent
Output:
Person A is stopping constantly on the way to Platform 6 and is being interrupted by something.
Input:
Person B is looking at a sign and around at the space before entering a lobby.
Non-Locomotor
Upper Body
Upper Body
Head Rotation
Output:
Person standing and looking around.
Passive
Looking for entrance
Quality
Fast
Output:
Person B is standing and looking around at a fast pace, looking for entrance to the building, possibly confused or in a rush.

Figure 3. Framework applied to example scenarios

3. Case Study

3.1. METHODOLOGY

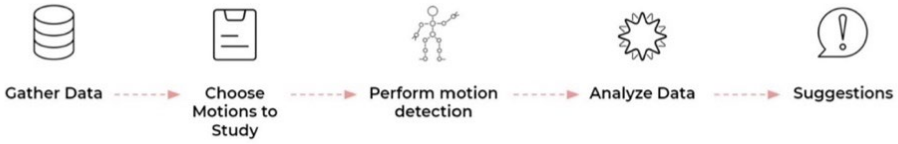

Figure 4. Methodology for Case Study

The goal of this case study is to develop an application of the proposed framework to initially assess its validity and applicability. Figure 4 shows the method process of the case study leading to a possible use case of the data which contributes to design suggestions. While all four taxonomy groups are discussed and proposed by the framework, this case study will focus primarily on State and Type, leading to Computer Identification. Inspired by the domestic motion studies done by Callaghan and Palmer (1944), household kitchens were selected as the target area of study. Two sets of video footage of kitchen usage by their inhabitants were collected. The first case, Kitchen 1, is a 25-minute recording of a 170 cm adult male making breakfast; the second case, Kitchen 2, is a 20-minute recording of a 163 cm adult female washing dishes.

Video footage was used for the purpose of obtaining a general understanding of the activities in the kitchen. In order to focus the scope of the case study, only EMSI *State* and *Type* groups were tested. To identify these, the footage was broken down into frames, a technique most closely related to the single-layered sequential approach (Aggarwal and Ryoo, 2011, p.14). Joint locations in each frame were detected using PoseNet (Papandreou et al., 2018), a machine learning model trained to estimate human skeleton poses. A custom algorithm was then developed to detect and measure the differences between frames to approximate a motion.

The algorithm detects for the following motions: translation, fixed, rotation, bending and reaching. Horizontal-Translation and Horizontal-Fixed *Types* are considered as a binary set for calculation. To determine the fixed or translating state, the average location of points detected by the pose estimation framework to be above the hip are computed. The average position of pose estimation in each frame is subtracted from the average position of the frame prior to it. If the change is greater than a predetermined value, the frame is temporarily noted as a translational motion. To reduce detection errors, the final assignment of motions is determined based on the current and two previous detections of motion, which can be seen in Figure 5.

Rotation, Bending, Reaching are all identified with JavaScript code based on the distance and positioning changes between shoulder, nose, and wrist locations over the series of video frames, with each detection requiring a manually adjusted threshold factor to refine accuracy.

3.2. RESULTS

After detecting the motions in the footage, a file containing information of the motions detected, time of detection, and motion coordinates were generated for data analysis. Figure 5 shows an example of the data generated from the video motion analysis.

Type:			Type:			Type:		
Translation	Time	X,Y	Rotation	Time	X, Y	Reaching/Bending	Time	X, Y
translating	9.54	30.82, 292.85	turning	9.52	30.82, 292.85	reaching	9.52	43.14, 312.76
translating	9.56	30.82, 292.85	no turning	9.53	30.82, 292.85	reaching	9.53	43.14, 312.76
fixed	9.58	30.95, 292.21	no turning	9.54	30.82, 292.85	reaching	9.54	43.14, 312.76
fixed	9.59	30.80, 291.67	no turning	9.56	30.82, 292.85	reaching	9.56	43.14, 312.76
…	…	…	….	…	…	…	…	…

Figure 5. Example of first few lines of the exported file

A series of motion density distribution graphs, shown in Figure 6, were generated to identify the types and locations of motion occurrences within the context of the environment. In kitchen 1, the graphs indicate the user mostly occupied the L-shaped area at the back of the kitchen and stands between the stove and the sink at the back. Rotational motions occur throughout the active space, while Bending only occurs between the stove and sink, and Reaching occurs near the cabinets on the left. In kitchen 2, motions are mostly concentrated in front of the sink. There is a fair amount of rotation and bending motion in front of the sink. There are also some occurrences of Reaching motion near the top of the fridge.

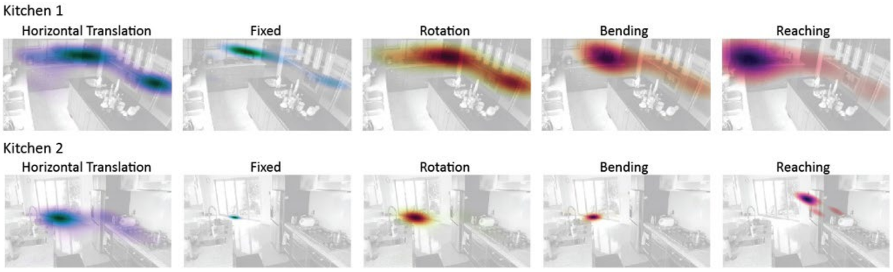

Figure 6. Graphs showing density and distribution of varying motion types in each kitchen space based on the activity recorded - Horizontal, Fixed, Rotation, Bending, Reaching

Design outcomes and suggestions can accommodate different action priorities and can change with varying case studies. For this work-in-progress case study, the data gathered is used to suggest areas of architectural rehabilitation to reduce reaching and/or reduce bending for users who have difficulties. Possible design suggestion areas to improve the user experience of the space in these two kitchens is to reduce bending and reaching motions by eliminating items stored in high or extremely low locations to avoid straining and inaccessibility. Figure 7 shows areas identified for possible architectural design adjustments in both kitchens.

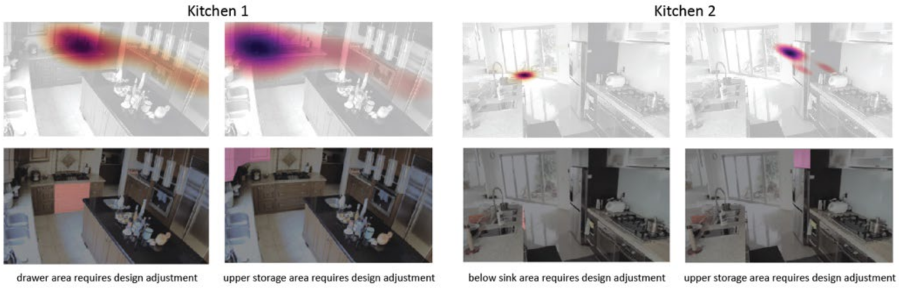
Kitchen 1
Kitchen 2
drawer area requires design adjustment
upper storage area requires design adjustment
below sink area requires design adjustment
upper storage area requires design adjustment

Figure 7. Suggested architectural design areas based on information from motion graphs

4. Discussion

4.1. LIMITATIONS

The framework is still under development, and further progress will address current limitations on the technical side. Factors such as lighting, and camera angles play a large role in the accuracy of the data. Experiments with two or more cameras located in different places in the same room would correct this limitation and help to provide more precise location data and motion information, particularly if one camera has any obstructed views from any furniture.

Potential privacy intrusions or ethical boundaries were considered throughout the research. While recorded videos with participants were used in the documented case study, unique private information related to participant identity, location and time is not important nor needed to conduct studies. The only information that is crucial is joint identification, which is conducive to anonymity, and highly encouraged in any further work. Video capture is not needed in the real-time data collection process.

Future case studies would benefit from incorporating object and environment recognition, as well as more refined motion detection models, to further draw relationships between physical context and motions identified.

Furthermore, the framework could also be applied to building simulations if the input is changed from real-time video to a series of data points or recordings proposed by agent-based pedestrian simulation software.

While the framework considers motion specific to architectural interaction, it does not go into depth regarding every listed category. It is merely proposed as a starting point for which to think about how interaction can be measured within different spatial contexts and what categories would be important to study. It has the potential to be used for studying movement within fields of accessibility, architectural rehabilitation, simulation, renovation, as well as broader fields of health and wellbeing that are crucial to the UN Sustainable Development Goals, such as neuroimmunology and influence on homeostatic conditions, psychology, universal design, and usability studies for the built environment (Dougherty et al. 2018). The study indicates research opportunities for pose and motion estimation in aiding spatial evaluation.

4.2. CONCLUSION

The Elemental Motion in Spatial Interaction (ESMI) framework of using typologies of motion to analyse spatial experience in architecture proposes a new approach to evaluating an architectural space by breaking down human movement in an occupied space. This approach reduces inherent biases when studying existing spaces, providing insight into how architectural space is used and experienced by differing individuals, depending on the motion efforts exerted.

The case study demonstrates one possible workflow of implementing selected parts of the framework, using the aid of existing pose detection algorithms. This workflow is at an early stage with several limitations and many possibilities for improvement. Further development will include performing analysis on more than one person at a time, with an enhanced algorithm to remove lag and improve accuracy. This can be further developed to detect architectural elements in a video to establish a measurable relationship between motion and environment. While the current iteration of the framework has four predominant contextual categories, there are many more factors that could be considered and tested to determine relevance, reliability, and utility.

References

Aggarwal, J. K., & Ryoo, M. S. (2011). Human Activity Analysis: A Review. *ACM Computing Surveys*, 43(3). https://doi.org/10.1145/1922649.1922653

Ajili, I., Mallem, M., & Didier, J.-Y. (2017, September). Robust human action recognition system using Laban Movement Analysis. *Procedia Computer Science*. 21st International Conference on Knowledge-Based and Intelligent I, Marseille, France. https://doi.org/10.1016/j.procs.2017.08.168

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2*, 1395-1402 Vol. 2. https://doi.org/10.1109/ICCV.2005.28

Bobick, A. (1997). Movement, Activity and Action: The Role of Knowledge in the Perception of Motion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 352*, 1257–1265. https://doi.org/10.1098/rstb.1997.0108

Bortolini, M., Gamberi, M., Pilati, F., & Regattieri, A. (2018). Automatic assessment of the ergonomic risk for manual manufacturing and assembly activities through optical motion capture technology. *Procedia CIRP, 72*, 81–86. https://doi.org/10.1016/j.procir.2018.03.198

Callaghan, J., & Palmer, C. (1944). *Measuring space and motion*. John B. Pierce Foundation.

Dougherty, B., & Arbib, M. (2013). The evolution of neuroscience for architecture: Introducing the special issue. *Intelligent Buildings International, 5*. https://doi.org/10.1080/17508975.2013.818763

Fox, M., & Polancic, A. (2012). Conventions of Control: A Catalog of Gestures for Remotely Interacting with Dynamic Architectural Space. *ACADIA 12: Synthetic Digital Ecologies*.

Gilbreth, F., & Gilbreth, L. (1917). *Applied motion study: A collection of papers on the efficient method to industrial preparedness*. Macmillan.

Hirschberg, U., Sayegh, A., & Zedlacher, S. (2006). 3D motion tracking in architecture: Turning movement into form—Emerging uses of a new technology. *Communicating Space(s) 24th ECAADe Conference Proceedings*.

Hu, Z., & Park, J. H. (2017). HalO [Indoor Positioning Mobile Platform]. *ACADIA 2017: DISCIPLINES & DISRUPTION [Proceedings of the 37th Annual Conference of the Association for Computer Aided Design in Architecture]*, 284–291.

Marey, E.-J., & Pritchard, E. (1895). *Movement*. D.Appleton.

Moeslund, T., & Granum, E. (2001). A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding,* 81, 231–268. https://doi.org/10.1006/cviu.2000.0897

Mohamed, A. (2015). A Novice Guide towards Human Motion Analysis and Understanding. *CoRR*. http://arxiv.org/abs/1509.01074

Muybridge, E. (1901). *The human figure in motion*. Chapman & Hall.

Nagymate, G., & Kiss, R. (2018). Application of OptiTrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronics, 5*. https://doi.org/10.17667/riim.2018.1/13

Newlove, J. (1993*). Laban for Actors and Dancers: Putting Laban's Movement Theory into Practice—A step-by-step guide*. Routledge.

Papandreou, G., Zhu, T., Chen, L.-C., Tompson, J., & Murphy, K. (2018). *PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model*. Retrieved from http://arxiv.org/abs/1803.08225

Peng, W., Zhang, F., & Nagakura, T. (2017). Machines' Perception of Space: Employing 3D Isovist Methods and a Convolutional Neural Network in Architectural Space Classification. *ACADIA 2017: DISCIPLINES & DISRUPTION [Proceedings of the 37th Annual Conference of the Association for Computer Aided Design in Architecture]*, 474–481.

Prinsloo, T.-T., Munro, M., & Broodryk, C. (2019). The Efficacy of Laban Movement Analysis as a Framework for Observing and Analysing Space in 'Rosas danst Rosas'. *Research in Dance Education,* 20, 331–344.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, 3*, 32-36 Vol.3. https://doi.org/10.1109/ICPR.2004.1334462