UNIVERSITY OF CALIFORNIA

Los Angeles

Causal Action:

A Framework to Connect Action Perception and Understanding

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Yujia Peng

2019

ABSTRACT OF THE DISSERTATION

Causal Action:

A Framework to Connect Action Perception and Understanding

by

Yujia Peng

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2019

Professor Hongjing Lu, Chair

Human actions are more than mere body movements. In contrast to dynamic events involving inanimate objects, human actions have a special status in that they control interactions with the world and afford privileged access to the experience of agency and to control interactions with the world. Several causal constraints on human actions support the generation and the understanding of actions. For example, human actions inherently involve a causal structure: limb movements generally cause changes in body position along a path through the environment to achieve intentional goals. However, it remains unclear how the system that supports action perception communicates with high-level reasoning system to recognize actions, and more importantly, to achieve a deeper understanding of observed actions. My dissertation aims to determine whether causality imposes critical motion constraints on action perception and understanding, and how causal relations involved in actions impact behavioral judgments. The project also investigates the developmental trajectory and neural substrate of action processing, and whether a feedforward deep learning model is able to learn causal relations solely from visual observations of human actions. Through behavioral experiments, an infant eye movement study, a neural study using magnetoencephalography, and model simulations, my dissertation yields a number of insights. 1) Humans implicitly and automatically rely on causal expectations to explain motion information when perceiving body movements and meaningful social interactions; 2) Sensitivity to causal constraints on actions develops early in infants even before 18 months of age, and shows a clear relation with the development of gross motor functions. 3) Congruency to causal relations involved in human actions can be decoded from neural MEG signals, with the processing of causal actions eliciting a distributed neural network. The brain network involves temporal, parietal, and frontal regions that are important locus for spatial relation reasoning, decision making, and intention understanding. 4) Recent stimulus-driven deep learning neural networks are unable to learn the causal relations involved in actions, failing to create high-level representations for causal actions. Overall, my dissertation reveals the importance of causality in bridging action perception and understanding, highlights the key missing computational components in deep learning models for complex visual stimuli such as human actions, and provides a potential framework to connect seeing and thinking.

The dissertation of Yujia Peng is approved.

Chris Hak Wan Lau

Scott Pratt Johnson

Yingnian Wu

Hongjing Lu, Committee Chair

University of California, Los Angeles

2019

Dedication

This dissertation is dedicated to my inspiring parents,

for their endless love, support, and wisdom;

My wonderful husband, Jiming,

for his love and knowledge that encourage me to pursue my dreams.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2 of this dissertation is a version of Peng, Thurman, and Lu (2017). Lu and I developed the study concept and all the authors contributed to the study design. The manuscript was written with theoretical contributions from Thurman and Lu. Chapter 3 is a version of Peng, Ichien, and Lu (under review). Lu and I developed the study concept and study design. Ichien and I developed the experimental stimuli. Both Ichien and Lu contributed to the theoretical components of the manuscript. Chapter 4 of this dissertation is a version of Peng, Nguyen, Brady, Lu, and Johnson (under review). Lu, Johnson and I developed the study concept and study design. Nguyen and I programmed the study. Shannon, Nguyen, and I performed data collection. Nguyen and I performed data analysis. All contributed to the theoretical components of the introduction and discussion of the manuscript. In chapter 5, Lu and I developed the study concept. Fang, Lu helped to develop the paradigms of the experiments. Gong and Chen contributed to data collection. Cushing and Lu contributed to data analysis. In Chapter 6 of this dissertation, Lu and I developed the study concept. Shu and I conducted the model setup.

It is a genuine pleasure to express my deep sense of gratitude to those who have supported me throughout my Ph.D. program. To my advisor, Dr. Hongjing Lu, your timely advice, meticulous scrutiny, scholarly wisdom, and scientific approach have helped me to a very great extent to achieve where I am now. You are a role model on many levels, and I am deeply grateful for your effort, patience, and encouragements in helping me to improve and succeed. To my other faculty mentors and collaborators, including but not limited to Drs. Brian Keane, Fang Fang, Frank Pollick, Hakwan Lau, Keith Holyoak, Marco Iacoboni, Martin Monti, Scott

Johnson, Song-Chun Zhu, and Yingnian Wu, I really enjoyed working with you and your support and input have been invaluable to my research.

To my collaborators, including but not limited to Drs. Gennady Erlikhman, Jeffery Chiang, Jeroen van Boxtel, Joseph Burling, Junzhu Su, Steven Thurman, Tawny Tsang, Tianmin Shu; Bryan Nguyen, Cody Cushing, Hannah Lee, Nicholas Ichien, Lifeng Fan, Marissa Ogren, Pria Daniel, Shannon Brady, and Xizi Gong. Thank you for giving me the opportunity to conduct research with such amazing people and for always providing valuable input to my research. To all of my RAs, it has been a pleasure performing research with all of you over the past five years.

To my parents and all other people in the family, you have always been there for me when I needed you. The distance never felt far even though we are separated by an ocean. To my beloved husband, Jiming Sheng, I'm so grateful to have you in my life. Thank you for making every day of my life shining brightly, for being my best friend, and for being a caring and trustworthy partner! Time has only deepened my love for you.

To all of my friends, including but not limited to Drs. James Kubricht, Melissa DeWolf; Akila Kadambi, JD Knotts, Junho Lee, Maureen Gray, Micah Johnson, Nicholas Baker, Sha Li, Yixin Zhu, and Zheng Kang, you have always helped me to gain valuable insights and to smile. It has been great having you along for the ride.

Portions of this work were supported by a CSC scholarship, UCLA Graduate Division Dissertation Year Fellowship, and UCLA GSRM awards.

VITA

EDUCATION

2014-2016	Masters of Arts, Psychology University of California, Los Angeles (UCLA)
2010-2014	Bachelor of Science, Psychology Peking University (PKU), Beijing, China

PUBLICATIONS

Peer-reviewed Journal Articles

Tsang, T., Ogren, M., Peng, Y., Nguyen, B., Johnson, K.L. & Johnson S.P. (2018). Infant Perception of Sex Differences in Biological Motion Displays. *Journal of Experimental Child Psychology, 173,* 338–350.

Keane, B. P.[*], Peng, Y.[*], Demmin, D., Silverstein, S. M., & Lu, L. (2018). Intact Perception of Coherent Motion, Dynamic Rigid Form, and Biological Motion in Chronic Schizophrenia. *Psychiatry Research, 268,* 53-59.

Shu, T.[*], Peng, Y.[*], Fan, L., Zhu, S., & Lu, H. (2017). Perception of Human Interaction Based on Motion Trajectories: from Aerial Videos to Decontextualized Animations. *Topics in Cognitive Science, 10*(1), 225-241.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal Action: A Fundamental Constraint on Perception and Inference with Body Movements. *Psychological Science, 28*(6), 798-807.

van Boxtel, J., Peng, Y., Su, J., & Lu, H. (2016). Individual differences in high-level biological motion tasks correlate with autistic traits. *Vision Research, 141,* 136-144.

Chen, J., Yu, Q., Zhu, Z., Peng, Y., & Fang, F. (2016). Spatial summation revealed in the earliest visual evoked component C1 and the effect of attention on its linearity. *Journal of Neurophysiology, 115*(1), 500-509.

Chen, J., He, Y., Zhu, Z., Zhou, T., Peng, Y., Zhang, X., & Fang, F. (2014). Attention-dependent early cortical suppression contributes to crowding. *The Journal of Neuroscience, 34*(32), 10465-10474.

Lu, J. *, & Peng, Y.* (2014). Brain-Computer Interface for Cyberpsychology: Components, Methods, and Applications. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 4*(1), 1-14.

Conference Proceedings

Peng, Y., Ichien, N., & Lu, L. (2019). Perception of Continuous Movements from Causal Actions. *Proceedings of the 41th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Shu, T., Peng, Y., Lu, H., & Zhu, S. (2019). Partitioning the Perception of Physical and Social Events Within a Unified Psychological Space. *Proceedings of the 41th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

[*] The first two authors contributed equally.

Peng, Y., Javangula, P. R., Lu, L., & Holyoak, K. J. (2018). Behavioral Oscillations in Verification of Relational Role Bindings. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Shu, T.*, Peng, Y.*, Fan, L., Zhu, S., & Lu, H. (2017). Inferring Human Interaction from Motion Trajectories in Aerial Videos. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, UK: Cognitive Science Society. (The work was awarded the *Computational Modeling Prize in Perception and Action* from the 39th Annual Meeting of the Cognitive Science Society, 2017 at London.)

Peng, Y., Thurman, S. & Lu, H. (2016). Causal action: a fundamental constraint on perception of bodily movements. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*. Philadelphia, Pennsylvania: Cognitive Science Society.

Book Chapter

Peng, Y. (2019). Brain-Computer Interface for Cyberpsychology. In *Analyzing Human Behavior in Cyberspace* (pp. 102-122). IGI Global.

PRESENTATIONS

Peng, Y., Javangula, P. R., Lu, L., & Holyoak, K. J. (2018, July). *Behavioral Oscillations in Relational Representations of Visually-presented Actions*. Poster presented at the 40th Annual Conference of the Cognitive Science Society. Madison, Wisconsin.

Peng, Y., & Lu, H. (2018, May 18-23). *Behavioral oscillations reveal hierarchical representation of biological motion.* Poster presented at the Vision Sciences Society, St. Pete Beach, FL.

Shu, T. *, Peng, Y. *, Fan, L., Zhu, S., & Lu, H. (2017, July 26-29). *Inferring Human Interaction from Motion Trajectories in Aerial Videos*. Talk presented at the 39th Annual Conference of the Cognitive Science Society. London, UK.

Peng, Y., & Lu, H. (2017, May 19-24). *Seeing illusory body movements in human causal interactions.* Poster presented at the Vision Sciences Society, St. Pete Beach, FL.

Peng, Y., Thurman, S. & Lu, H. (2016, August 10-13). *Causal Action: A Fundamental Constraint on Perception of Bodily Movements.* Poster presented at the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA.

Peng, Y., Thurman, S. & Lu, H. (2016, May 13-18). *Phenomenal Causality in Biological Motion Perception.* Talk presented at the Vision Sciences Society, St. Pete Beach, FL.

FUNDINGS and AWARDS

Dissertation Year Fellowship, 2018-19, UCLA

Computational Modeling Prize in Perception and Action, the 39th Annual Meeting of the Cognitive Science Society, 2017, London

Liu Yunghuo Bei Qui Memorial Fellowship, 2017, Department of Psychology Endowed Fellowships, UCLA

Graduate Summer Research Mentorship award, 2015 and 2016, UCLA

Young Scientist Travel Awards, 2016 annual meeting of the Cognitive Science Society

China Scholarship Council (CSC) -UCLA Stipend Scholarship

CHAPTER 1

Introduction

Recognizing human body movements is considered as one of the most sophisticated abilities supported by the human visual system. As suggested by Darwin (1872), “actions speak louder than pictures when it comes to understanding what others are doing and feeling”, human actions readily communicate information such as goals and intentions. With a glimpse, we can extract various information from even highly degraded motions such as point-light displays. Since the early work of Gunnar Johansson (1973), point-light displays have been widely used to study how people perceive biological motion from only a few moving joints. Human adults have been shown to be proficient in perceiving sophisticated social-relevant information from point-light displays, such as gender (Kozlowski & Cutting 1977; Mather & Murdoch 1994), identity (Cutting, & Kozlowski, 1977; Pavlova, 2011), personalities (Brownlow et al. 1997; Montepare & Zebrowitz-McArthur 1988; Gunns et al.2002), emotions (Dittrich, Troscianko, Lea, & Morgan, 1996; Clarke, Bradshaw, Field, Hampson, & Rose, 2005) and intentions (Runeson & Frykholm 1983). However, the interplay between perception and thinking has been understudied in the field, specifically, the connection between action recognition and reasoning about actions remains unclear.

Human actions are more than mere body movements. In contrast to dynamic events involving inanimate objects, human actions have a special status in that they afford privileged access to the experience of agency and to control interactions with the world. Through interventional actions, humans are equipped to discover causal relations in the physical and social environment (Abravanel, Levan-Goldschmidt, & Stevenson, 1976; White, 1999). Even without actively manipulating objects through goal-directed interventions, navigating one's body through the environment provides direct experience of cause-effect relations. In his *Philosophical Investigations*, Ludwig Wittgenstein (1953, p. 161) posed a famous question: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?" This question highlights the special status of the human actions relative to other dynamic events in the natural world, that action is a combination of willing plus movement (Jaeger, 1973). In other words, human actions are usually initiated by the underlying goals and desires and often involve processes that *cause* willful bodily movements. The human body navigates the environment via locomotory movements that leverage gravity and limb biomechanics to propel the body in a particular direction. This process creates a causal link between limb movements and whole-body translation, resulting in expectations about the relation between the two motion cues (i.e., relative limb movements with reference to body-centered coordinates as related to body displacements with reference to distal world coordinates). "To step up by lifting one's leg" is properly considered to include the objective content of action: it involves a causal structure in which moving one's limbs in a certain way provides a means to cause position changes of the body such that the intentional goal of stepping up is fulfilled. Actions thus afford privileged access to experiencing the role of agency and provide a powerful tool to produce interventions that in turn help to discover causal relations in the physical and social environment (White, 1999).

Critically, causal constraints of human actions support the generation and the understanding of actions. For example, the sense of agency facilitates the inference to identify the cause of an observed action. We often have a strong sense of causality as actions unfold. For example, observing certain limb movements triggers the expectation of changes in body position.

In the context of social interaction, one person's actions may serve as the cause to make a second person acts in a certain way. The ability to understand the actions of others is closely related to our own internal sensorimotor representations (Decety & Grèzes, 1999), and would also seem to involve access to ingrained representations of these inherent causal dynamics. Despite the importance of causal action, few studies have looked into how people perceive causal action with whole-body movements. Thus, it remains unclear how the human visual system communicates with the high-level cognitive system to not only recognize actions, but more importantly, make inference about intentions behind observed actions. Understanding how causal actions are perceived and inferred will make a bridge between perception and reasoning and will provide important insights into the future development of artificial intelligence.

A few questions remain to be answered: (i) whether humans are sensitive to causal relations that are inherent in human actions; (ii) how does causality in human actions influence action perception; (iii) how does the ability to perceive and understand causal constraints in human actions develop in infants; (iv) what are the neural mechanisms underlying the causal understanding in human action perception; and (v) whether the ability of inferring causality from human actions can be captured by the state-of-art computational models. To address these questions, my dissertation develops a framework to systematically examine the connections between action perception and reasoning from behavioral, neural, and computational aspects. Chapter 1 provides an overview of research endeavors and a compressive review of the literature. Chapter 2 examines if human observers spontaneously use causal relations as a cue when perceiving human actions and if the violation of causality influence the perceived naturalness of human actions. Chapter 3 studies the top-down influences of causality on the perception of body movements when observing human interactions with physical objects and agents. Chapter 4 investigates the developmental trajectory of causality in the form of a motion consistency constraint, and how this constraint impacts the perception of human body movements in infancy and relates to motor developments. Chapter 5 investigates the neural mechanism underlying the perception of causal action using Magnetoencephalography (MEG) and provides evidence to unveil the spatiotemporal dynamics of neural activities when observing causal versus non-causal human actions. Chapter 6 examines if the state-of-art convolutional neural network (CNN) could account for human behavior in the causal perception of human actions.

Taken together, these lines of research aim to advance our understanding of the role of causality in human action perception and the potential mechanisms that support the visual processes of causal action perception. Specifically, a central goal of this dissertation is to point out that causality serves as one of the important motion constraints and modulates human action perception constantly and automatically. The close interaction between bottom-up processes of motion signals and top-down processes of causal reasoning together support our sophisticated processes of action understanding and goal inference. Evidence from the current work calls for attention on bringing high-level causal reasoning into the research of human action recognition and the development of artificial intelligence.

The following sections provide motivations and backgrounds for the studies reported in the chapters herein. Section 1.1 discusses the role of causality in human perception as found in previous studies using mainly physical movements or simple dynamic events, showing that causality facilitates the temporal binding and spatial binding of causes and effects. To build on this existing knowledge, Chapter 2 and 3 will extend these studies to action perception by showing the evidence that causality strongly influences the visual perception and understanding of human actions. The corresponding experiments in Chapter 2 and 3 are the first to study the role of causality on human action perception and social cognition. Section 1.2 reviews the previous evidence of the emergence of the causal understanding in infancy, showing a tight relation between causal perception, action understanding, and motor functions. Chapter 4 tests the development of the perception of causal constraints in human body movements for 9- to 18-month infants and examines if the development of causal perception is tied to the development of gross motor functions. Section 1.3 outlines the previous neural evidence underlying perceptual causality and Chapter 5 presents the first evidence of the neural representation of the causal constraint in human actions using MEG. Section 1.4 provides a brief overview of computational modeling works on human action perception and Chapter 6 tests the performance of the state-of-art two-stream CNN on recognizing causal versus non-causal human actions. Finally, Chapter 7 summarizes and discusses the findings through chapters 2 to 6 and proposes future directions.

1.1 The Phenomenal Causality in Moving Objects

We live in a dynamic world with various objects movements, from simple Newtonian motion of inanimate objects to complex human motion. Motion provides us information far beyond simple physical displacement, but also important high-level information such as the causal relation between moving objects. The ability to correctly and efficiently understand the causality in motion displays plays an important role in our interaction with others and the distal world. Numerous evidence has demonstrated an intimate connection between causation and perception (Michotte, 1963; Heider & Simmel, 1944; Scholl & Tremoulet, 2000; White, 2006). Dating back to Michotte's seminar work in 1963, the launching effect demonstrated that the collision of two balls led to a strong perception of causality. In the classical visual display of the launching effect, observers view a moving object, referred to as the launcher, that contacts a stationary object, referred to as the target. Immediately after the contact, the target object begins to move and the launcher stops. Even though many possible interpretations (e.g. the launcher was stopped by the target or by friction) can be derived from the visual input, observers consistently attribute motion of the target to the launcher, or that the motion of the launcher causes the movement of the target. The perception of causality occurs almost automatically and is consistent across subjects. Hence, the launching effect opens a window for researchers to investigate human's causal perception given its straightforward motion input and the associated robust responses.

In considering the role of causality in perception, a particularly important theory is the temporal priority principle (Hume, 1739/1888) that causes always come before their effects in time, even though there may appear to be some circumstances in which cause and effect appear perfectly contemporaneous (e.g. pressing a button of a TV remote to turn the volume up). The temporal relationships between actions and subsequent outcomes determine how causal relations are inferred (Shanks, Pearson, & Dickinson, 1989). For instance, temporally contiguous outcomes are more likely to be perceived as generated by our actions (Farrer, Valentin, & Hupé, 2013, Wegner and Wheatley, 1999, Young, 1995). Causal beliefs have been known to distort people's perception of time, biasing people to perceive causally-related events closer in time than unrelated events (Faro, Leclerc, & Hastie, 2005; Buehner & Humphreys, 2009). This distortion is consistent with Hume's (1739/1888) principles of causality that causally-related events are spatially and temporally contiguous. Relevant body of prior research showed the effect of intentional binding (Haggard, Clark, & Kalogera, 2002), which have shown that the perceived time elapse between an intentional action (e.g., a key press) and its subsequent sensory effect (e.g., a tone or flash) is compressed, such that all sensory events following an action appear to draw closer in time to that action. Buehner (2012) showed that this type of temporal binding effect results from a general causal relation linking actions to their consequences.

Most previous studies examined the causality in actions using bimodal stimuli with a motor act as a cause and visual/auditory stimulus as the effect (Buehner & Humphreys, 2009; Faro, Leclerc, & Hastie, 2005; Shanks, Pearson, & Dickinson, 1989; Buehner, 2012; Haggard, Clark, & Kalogeras, 2002). Because bimodal stimuli recruit processes of sensorimotor integration, temporal asymmetry effects observed with such stimuli may result from other non-causal mechanisms (Stetson, Cui, Montague, & Eagleman, 2006; Rohde, Scheller, & Ernst, 2014). Here, I propose to test causality in human actions as unimodal stimuli. When we consider the relation of limb movements and body motion, the contiguity constraint is usually satisfied given that both movements arise from a single agent. Although previous research has shown that action recognition can be influenced by the temporal congruency between limb movements and body displacements (e.g., Masselink & Lappe, 2015; Thurman & Lu, 2013, 2016a), no direct evidence has established that humans rely on the directionality of the cause-effect relation between the two motion cues in perceiving and making inferences about human actions. To address this issue, Chapter 2 and 3 present behavioral evidence of the important role of causality in guiding the perception of human actions. Specifically, Chapter 2 tested how causality influences the perceive naturalness and the understanding of human body motions using simplified motion displays. Chapter 3 examined how causality shapes our perception of inter-human social interaction in natural scenes.

1.2 The developmental trajectory of causal perception

While studies from human adults provide important evidence on the role of causality in human action perception and understanding, the origin of causal action perception remains unknown. Specifically, it remains unclear how causal action perception emerges in the developmental trajectory and what factors influence the development. Unlike adults, whose behavior and judgments are based on variance strategies and prior knowledge, behaviors of infants provide direct evidence on how certain stimulus features influence visual processes of action perception and understanding. In addition, testing infants at different stages of developmental trajectory could answer when certain traits emerge and how it interacts with the development of other functions such as motor abilities. Hence, developmental studies with human infants are necessary to provide answers to when and how causal action perception emerges in human beings.

Learning the causal relations in everyday action sequences is no easy task for infants. Infants need to break a continuous action sequence into meaningful subsequences and infer which action leads to the outcome (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011). One of the central questions is whether causal perception is innate or it develops through learning from life experience. Furthermore, even if infants are able to discriminate causal events and non-causal events in the typical paradigm of causal perception, it’s debatable that whether it’s based on visual spatiotemporal features or on abstract causal representation. To investigate the emergence of causal perception, previous studies focused mainly on physical motion sequences such as the launching events (Michotte, 1963). It seems that the learning of causality started from a very early stage of infancy. Oakes & Cohen (1990) found that starting at roughly 10-month, infants start to discriminate launching events on the basis of causality. Other evidence showed that the development of causal perception might origin as early as 6-month (Leslie, & Keeble, 1987) even though the causal perception was possibly mainly driven by the spatiotemporal features embedded in the perceived dynamic scene (Leslie, 1982). However, it cannot be guaranteed that the findings can be generalized to human actions given that human actions are different from physical object movements in the ways that human actions involve self-initiated movements and are driven by beliefs and goals. For infants, the learning of causality in human actions cannot be solely driven by visual spatiotemporal features but require a deeper understanding of the causality in actions. Other studies involving older kids have directly tested children's use of the temporal priority principle and evidence suggests that children as young as three years old use this principle to guide causal reasoning. For example, 3-year-old children more frequently chose the event that happened earlier in time as the cause (Bullock and Gelman, 1979), and 3- to 4-year-old kids were able to answer causal questions based on the temporal order of actions and consequences (Kun, 1978; Rankin & McCormack, 2013).

The development of causal perception has a close connection to the development of action perception and understanding. It has been found that 5-month-old infants showed selective attention to the goal of actions when an actor demonstrated grasping toward toys (Woodward, 1998). By 10- to 11-month, infants showed a novelty preference for disrupted causal action sequences (e.g. paused in the midst of an actor reaching for the towel), suggesting an early understanding of goal-oriented actions (Baldwin, Baird, Saylor & Clark, 2001). By 12-month-old, evidence has shown that infants interpret attentional behaviors (e.g. goal-directed points and eye gaze) in terms of the objects that they are directed towards (Woodward, 2003; Woodward & Guarjardo, 2002), suggesting the ability to predict the goal of observed actions. Taken together, by the end of the first year, infants already have basic knowledge of the goal-directed human actions.

The knowledge of the causal relations between one’s action and the goal also determines the producing of actions that can fulfill one’s own desire. Empirical evidence has shown that self-produced causal action may be necessary to promote the development of causal perception and the impact on the development of causal perception started as early as 4 ½-month-old (Rakison & Krogh, 2012). It was proposed that 9 to 12 months of age may mark a transitional time in infants’ ability to spontaneously generate goal-directed action sequences (e.g., Bates, Carlson-Luden & Bretherton, 1980). By 12 months of age, infants selectively imitate the goal of a sequence but often exclude the means of the sequence when they are not necessary for goal attainment (Carpenter, Call & Tomasello, 2005). By 18-month, infants were found to be able to reproduce other people’s actions based on the ultimate goal even when the observed actions failed to reach the goal (Meltzoff, 1995).

Given the close connection between causal constraints and action understanding, it is important to understand the development trajectory of the perception of causality in human actions. Through eye tracking experiments, Chapter 4 provides empirical evidence on the emergence of the perception of causal constraints in human actions and also on the relation between causal action perception and motor development. Results support that the understanding of causal constraints may appear as early as 18-month and the causal perception has an important connection to gross motor developments.

1.3 Neural Bases for Perceptual Causality and Action Recognition

In addition to using behavioral tests, neuroimaging research offers a complementary approach to building a more complete model of how causal constraints in actions are mentally represented. Neuroimaging studies provide evidence that cannot be answered by behavioral studies, such as the fine time course of the cognitive process even before reaching consciousness, a genuine reflection of representations of stimuli without the interference of motor functions, the ability to predict human behaviors based on decoding algorithms, and neurological knowledge that may benefit treatments of abnormal populations.

A vast number of studies have looked into the neural mechanism underlying the perceptual causality but were mostly focusing on simple causal events such as the "launching" effect. For example, in Fugelsang and his colleagues' work (2005), they used functional magnetic resonance imaging (fMRI) to examine the neural correlates of perceptual causality while participants were viewing alternating blocks of causal launching events and non-causal events with either a spatial or a temporal gap between the movements of the two objects. They found that causal events elicited significantly greater higher activation in the right middle frontal gyrus and in the right inferior parietal lobule than non-causal events, suggesting the important role of right hemisphere loci for perceiving causality for launching displays. Blos and his colleague (2012) further studied the neural processes involved in causal perception in different contexts: physical causality and social causality. Participants judged causal relationships of two objects in animated video clips which include either physical context or social context while providing matched contexts in terms of temporal and spatial characteristics. In the physical context, the blue ball yielded contact with the red ball. After a short delay, the red ball began to move in one of the seven directions. In the social context, the red ball was positioned above the center and the blue ball first moved horizontally and then advanced to one of the seven directions once it reached the middle, illustrating that the behavior of the blue ball was changed due to the presence of the red ball or the blue ball was attracted by the red ball and curved the trajectory. Results showed that the tasks in both contexts recruited similar brain areas, but elicited very different or even opposite activation patterns. For example, the activation of left insula increased for causal trials in physical contexts but not for causal social contexts, suggesting no significant activation for causal judged stimuli contrasted to non-causal judged stimuli. This indicates that the perception of causality is not a universal cognitive process that is implemented in certain brain regions, but, instead, depends on the context of events (i.e. physical or social). Similarly, Wood and his colleague (2014) investigated the neural systems involved in the elemental space perception, time perception and decision-making in the process of perceiving causality and showed that causal perception is not achieved by the activation of certain brain areas. Instead, depending on the context of the events and the spatiotemporal relation between objects, different brain areas are activated. Hence, a flexible brain network is recruited to process visual information in the way of supporting causal perception.

However, it still largely remains unknown regarding how we understand the goal and intentions embedded in human actions. One classic hypothesis "direct-matching hypothesis" states that we understand actions by mapping the visual representation of observed actions to our motor representation of performing the same action and the process is supported by the "mirror" neuron system, first found in monkey at ventral premotor cortex (Gallese, Fadiga, Fogassi,& Rizzolatti, 1996). The hypothesis was later supported by evidence of humans (Iacoboni, Woods, Brass, Bekkering, Mazziotta, & Rizzolatti, 1999; Nishitani, & Hari, 2000) showing that certain brain regions were more activated during the execution of actions on top of action observation, including Broca's area and the parietal lobe. The direct matching hypothesis provides a potential explanation for the mechanism behind action understanding. However, specific neural mechanisms of how the visual system extract relations between body movements and goals remains unclear.

One important step toward answering the question is to investigate the spatiotemporal dynamics of brain activities underlying the processing of causal constraints that support goal-orientated human actions. Using MEG, Chapter 5 will present neural evidence on the temporal dynamics of brain activities when watching human body movements that follow causal constraints and that do not follow causal constraints. The evidence fills a blank in the field of the neural basis of the visual perception of human actions and provides important insights into the high-level cognitive process of action perception.

1.4 Computational Models and Deep Learning Algorithms of Action Recognition

Given the behavioral and neuroimaging evidence as shown in previous chapters, it is clear that causality plays an important role in human action perception. However, the current field still lacks a computational model of human action perception that takes causality into consideration to address the phenomena as shown in human behavior. A successful computational model that has an understanding of causal relations between motion components is the foundation of understanding goals underlying human actions. This will greatly facilitate the development of artificial intelligence which can understand human behavior, interact with humans in a meaningful manner, and predict human action and desires efficiently.

To address the question of how people perceive actions, numerous studies have proposed computational models on the visual process of actions focusing on different visual constraints in actions. One representative model of action perception is by Giese and Poggio (2003), who developed a neurophysiologically plausible and quantitative model with two parallel processing streams: a ventral pathway and a dorsal pathway. The ventral pathway is specialized for the analysis of form in static frames. The dorsal pathway is specialized for optic-flow/motion information. Both pathways comprise hierarchies of neural feature detectors that extract form information or optic-flow information with increasing complexity as the hierarchy going up.

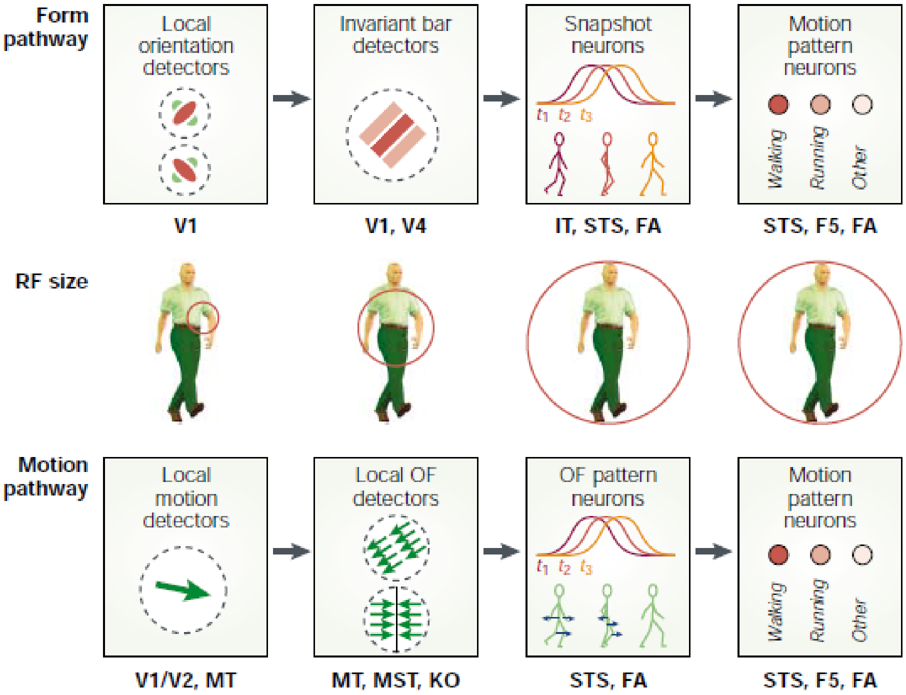

Figure 1.1: The hierarchical neural model of human action perception with two pathways by Giese & Poggio (2003). The top row and the bottom row represent the form and motion pathways. The middle row shows the size of receptive fields (RF) compared to a human figure. Insets and labels

show the different types of neural detectors at different levels of the hierarchy. IT, inferotemporal cortex; KO, kinetic occipital cortex; OF, optic flow; STS, superior temporal sulcus; V1, primary visual cortex.

Despite the fact that the model can address many neurophysiological findings, it is developed based on many experimenter constraints and models a simplified version of the complex visual processing of human actions. One of the drawbacks is that the model was built in a pure feedforward manner for both pathways, without top-down influence from other brain regions beyond visual areas. This makes the model efficient in perceiving motions based on low-level motion inputs but limits the capability under circumstances that require attention shifts and prior knowledge such as physical laws and biological structure of human bodies. Another limitation was that the model was not capable of recognizing dynamic events in natural scenes but only highly controlled biological motion stimuli.

In recent years, computer vision has developed a strong interest in hierarchical neural network architectures, or in other words "deep learning architectures" or "convolutional neural networks" (CNN) (LeCun, Bottou, Bengio, & Haffner, 1998). Such architectures outperformed other algorithms and even reached human-level performance on many visual tasks, first for image classification (Krizhevsky, Sutskever, & Hinton, 2012), and object recognition (LeCun, Bengio, & Hinton, 2015), but later also for many other computer vision problems including action classification (Karpathy, Toderici, Shetty, Leung, Sukthankar, & Fei-Fei, 2014, Le, Zou, Yeung, & Ng, 2011). It has been shown that deep learning architectures trained with proper datasets can develop units in layers whose tuning properties resemble the ones of neurons in areas V4 and IT (inferotemporal cortex) (Yamins, Hong, Cadieu, Solomon, Seibert, & DiCarlo, 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016).

One of the successful models using CNN was done by Simonyan and Zisserman (2014), who extended deep convolutional neural networks and developed two-stream CNNs which include the spatial and temporal pathways. A spatial CNN processes appearance information and a temporal CNN processes the optical flow information. At a later stage, these two information sources are integrated so that both appearance and motion information is used to recognize actions. The model performed well on action classification of two challenging datasets: UCF-101 which includes 13320 videos covering 101 action types (Soomro, Zamir, & Shah, 2012), and HMDB-51 which includes 6766 videos covering 51 action types (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011). The two-stream CNNs achieved the accuracy levels in the range of 55% to 70% for the HMDB-51 dataset (compared to a chance level of 2%). This performance represents a significant improvement over previous action recognition models based on deformable templates, or part-based approaches.

Despite the success of two-stream networks in recognizing actions, they share a common nature that the learning is based on a pure stimulus-driven manner and requires a large amount of training date, which is greatly different from the development of human minds. It remains unknown whether two-stream CNNs trained with natural videos yield representations of actions similar to those acquired by the human visual system, and whether CNNs perform similarly to humans in recognizing actions in novel displays that require an understanding of motion constraints such as causality. Chapter 6 examines the performance of two-stream CNNs on the perception of causal and non-causal actions and discusses the limitations and future directions of modeling human action perception.

CHAPTER 2

Causal Action: A Fundamental Constraint on Perception and Inference about Body Movements

2.1 Abstract

The human body navigates the environment via locomotory movements that leverage gravity and limb biomechanics to propel the body in a particular direction. This process creates a causal link between limb movements and whole-body translation. However, it is unknown whether humans use this causal relation as a constraint in perception and inference with body movements. In the present study, participants rated actions of other individuals as more natural when limb movements (as a cause) occurred before body displacements (as an effect) than when limb movements temporally lagged behind body displacements. This causal expectation for human body movements not only affected perceptual impressions regarding the naturalness of observed actions but also guided the interpretation of motion cues within a more generalized causal context. We interpret these results within a framework of causality as evidence that the constraint of causal action plays an important role in perception and inference with body movements.

2.2 Introduction

Seminal works (Heider & Simmel, 1944; Michotte, 1946/1963), coupled with contemporary developments (e.g., Scholl & Tremoulet, 2000; White, 2006), have demonstrated an intimate connection between causation and perception. A causal impression can directly arise from our perception of the world and can influence further perceptual judgments, such as event timing (Bechlivanidis & Lagnado, 2013, 2016). However, previous studies of causal perception have mostly focused on the interactions of inanimate objects in the physical world (e.g., colliding balls), which has limited generalization to more complex visual inputs, such as human actions. In contrast to dynamic events involving objects, human actions have a special status in that they afford privileged access to the experience of agency and enable discovery of causal relations in the physical and social environment through purposeful interventions (Abravanel, Levan-Goldschmidt, & Stevenson, 1976; White, 1999).

Consider one simple example of human actions. The human body navigates the environment via locomotory movements that leverage gravity and limb biomechanics to propel the body in a particular direction. This process creates a causal link between limb movements and whole-body translation, resulting in expectations about the relation between the two motion cues (i.e., limb movements in the body-centered reference frame in relation to body displacements in the environmental reference frame). This causal linkage may help explain why the “moonwalk” dance move popularized by Michael Jackson is experienced as surprising or even thrilling. While the dancer moves his or her legs in a way that appears to simulate walking forward, the whole body glides seamlessly backward, creating a dramatic conflict with the expected relationship between limb movements and body displacements.

Recent research has revealed that humans are sensitive to the temporal binding between limb movements and body displacements, given that we commonly observe the two types of motion occurring in near synchrony. Disrupting the temporal congruency between the two sources of motion information curtails the perception of animacy (Thurman & Lu, 2013), the detection of social interaction between two agents (Thurman & Lu, 2014a), and the discrimination of locomotion style (Masselink & Lappe, 2015; Thurman & Lu, 2016a). However, it is unclear whether people show tolerance to some situations in which limb movements and body displacements are temporally misaligned but in a causally consistent way (e.g., limb movements may be shifted ahead in time but still precede body displacements in locomotion). In other words, is the degree of tolerance constrained by the directionality of the causal relation between the motion cues?

The present study addresses this question by examining how the cause-effect relation inherent in body movements affects the perception and inference of actions. We used a key manipulation based on a ubiquitous feature of causation: the temporal-priority principle, which holds that a cause must precede its effect (Hume, 1739/1888; Price, 1992; White, 2006; Bechlivanidis & Lagnado, 2013). In Experiment 1, we systematically manipulated the direction and magnitude of temporal offsets between limb movements and whole-body displacements. If the causal relation between the two motion cues is important, the temporal-priority principle would predict that when body displacements (the effect) temporally lag behind limb movements (the cause), observers may show greater tolerance to a deviation from close simultaneity. In this situation, the temporal relation between limb movements and body displacements remains qualitatively consistent with normal causal directionality. In contrast, when body displacement (the effect) is shifted earlier in time to occur before the supposed cause (limb movements), observers may show less tolerance because of the strong violation of the causal expectation. However, if temporal alignment per se is the critical factor (without consideration of causal directionality), then we would expect a symmetric influence of temporal offsets on perceived naturalness of actions. In Experiment 2, we varied the cover story associated with identical stimuli to examine whether inference judgments shift to conform to the causal context when different beliefs are induced. Together with a series of control experiments, results from the present study provide evidence against mere associative learning and instead support the hypothesis that causal relations in body movements play an integral role in our perception and inference with actions.

2.3 Experiment 1

Experiment 1 was designed to assess how the directionality of temporal offsets between limb movements and body displacements affects the perceived naturalness of human actions.

2.3.1 Method

Participants. One hundred nine online participants were recruited through Amazon's Mechanical Turk (MTurk). Each was paid $1 for participating in the online experiment (average duration of 8 min). All experimental procedures were approved by the Committee for Protection of Human Subjects at the University of California, Los Angeles (UCLA). The sample size was determined on the basis of previous research on action recognition using MTurk (Shu, Thurman, Chen, Zhu, & Lu, 2016). Data collection for the online experiment stopped on the day when the expected sample size was reached.

Stimuli. Action stimuli were generated from the Carnegie Mellon University Motion Capture Database (http://mocap.cs.cmu.edu) and processed using the Biological Motion Toolbox developed by van Boxtel and Lu (2013). We selected actions in which a person walked on an uneven surface with invisible steps, and both horizontal and vertical body displacements were included in the action sequence. The stimuli used in the experiments appear in Videos S1 and S2 in the Supplemental Material available online; they can also be viewed at http://cvl.psych.ucla.edu/causal-action-2016.html.

Body displacements were computed as the change in the average position of the two hip joints in time, and limb movements were defined as the residual motion after subtracting the body-motion component on a frame-by-frame basis. The temporal relationship between limb movements and body displacements was manipulated by shifting the sequence of body displacements forward or backward in time relative to the sequence of limb movements, as illustrated in Figure 2.1a. Body displacements were manipulated to either lead or lag behind the posture change resulting from the limb movements. In the lead condition, the temporal sequence of body positions was shifted forward in time relative to limb movements (i.e., the effect preceded); in the lag condition, the temporal sequence of body positions was shifted backward in time (i.e., the cause preceded).

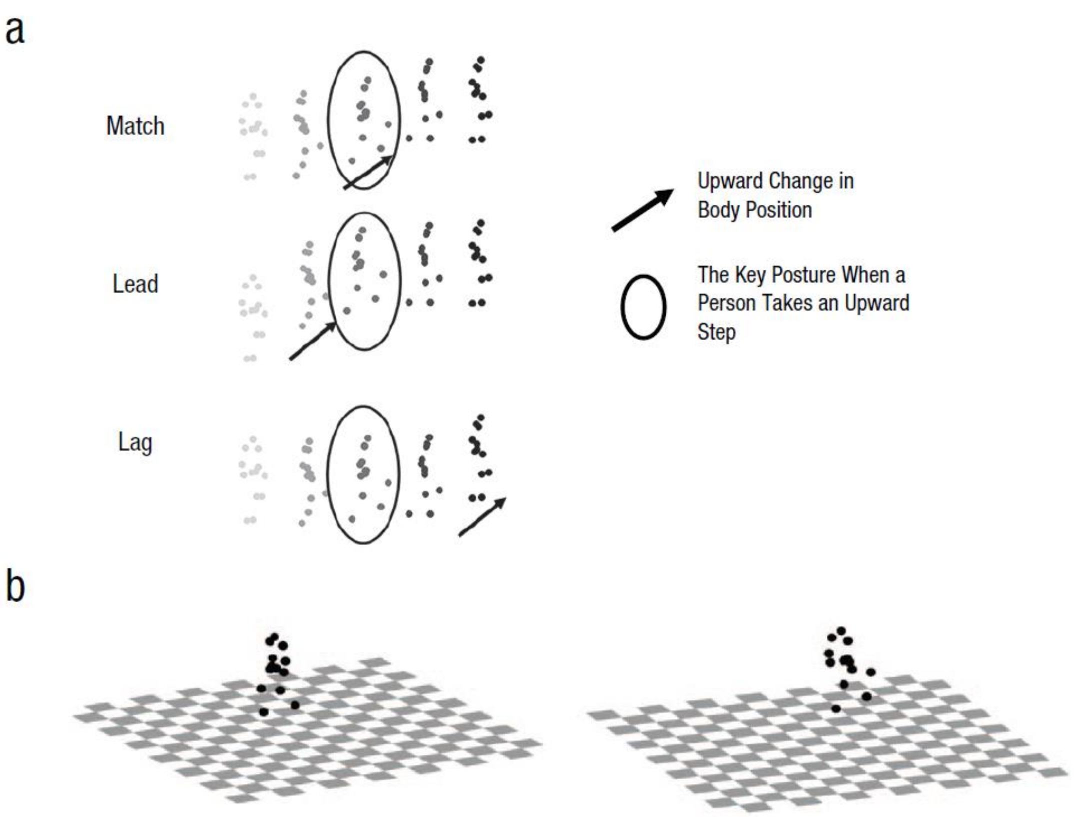
a
Match
Lead
Lag
Upward Change in Body Position
The Key Posture When a Person Takes an Upward Step
b

Figure 2.1: Illustrations of the stimuli in Experiment 1. The dots in (a) represent point-light walkers with different temporal relationships between body displacements and limb movements. The ellipses circle the key posture when a person takes an upward step, and the black arrows indicate the upward change of body position that is associated with such a step. The point lights in the walker changed from light to dark color to denote elapsed time. In the match condition, the posture and body displacements were in synchrony. In the lead condition, the body position changed before the limbs moved. In the lag condition, the body position changed after the limbs moved. The stimulus frames in (b) were taken from a sequence in which a point-light walker moved on an uneven surface that had invisible steps. The image slowly rotated in a clockwise direction. Videos S1 and S2 in the Supplemental Material show examples of dynamic stimuli. The videos are also available at http://cvl.psych.ucla.edu/causal-action-2016.html.

Procedure. Participants were presented with the following cover story: “Imagine you are viewing a walking sequence on an uneven surface with invisible steps through a slowly rotating camera. The rotation of the camera will help you perceive the 3D space. Look at the relative limb movements of the walker. Look at how the body position changes over time. Ask yourself if that could be a real person’s motion in the environment.” Participants were asked to rate the naturalness of the videos on a scale from 1 (unnatural) to 5 (natural).

On each trial, participants saw a point-light actor walking on a checkerboard surface with invisible steps, as shown in Figure 2.1b. The camera rotated in a counter-clockwise direction (meaning that the checkerboard and the actor appeared to rotate clockwise) at a speed of 3°/s, which was intended to facilitate 3-D perception of the action stimulus in the environment. Each video lasted 8.33 s and consisted of 500 frames selected from the original 2,000-frame videos based on the 32-s motion-capture data.

Actions were presented with a randomly selected starting viewpoint (45 from a side view) to ensure that any effect did not depend on a specific viewpoint. Six temporal offsets between limb movements and body displacements were used: 0, 0.5, 1.0, and 8.33 s (corresponding to 0, 30, 60, and 500 frames at a refresh rate of 60 Hz). The conditions with no offset and a large offset (i.e., 8.33 s) served as extreme cases to help participants anchor the two ends of the rating scale. Positive offsets constituted the lead condition (i.e., the effect of body displacements occurred before the causal cue of limb movements); negative offsets constituted the lag condition (i.e., the effect of body displacements occurred after the causal cue of limb movements).

The experimental procedure included two blocks. Each block consisted of 12 experimental trials (two starting viewpoints for each of six temporal offsets) and one randomly placed attention-check trial. The attention-check trial assigned a trivial task; participants were presented with either walking or jumping sequence and were asked to identify the presented action. The purpose of including these two attention-check trials was to identify outlier participants who gave random responses in the online experiment.

2.3.2 Results

Fourteen of the 109 online participants were removed from the analysis because they failed to satisfy the inclusion criteria. Specifically, 9 participants were excluded because they failed to recognize the simple actions in both of the attention-check trials. Data from 5 participants were excluded because they provided the same ratings for all trials in the experiment.

As expected, naturalness ratings were highest in the zero-offset condition (i.e., perfect synchrony between limb movements and body displacements; M = 3.82, SD = 0.79); naturalness ratings were lowest in the condition with a temporal offset of 8.3 s (M = 2.29, SD = 1.01). These results demonstrate that human observers are generally sensitive to the magnitude of temporal offsets between limb movements and body displacements. The two extreme conditions (i.e., 0.0 s and 8.3 s) did not include the directional temporal shifts to generate the lead and lag offsets; consequently, ratings for these conditions were not included in the following analyses.

To examine how the directionality of temporal offsets between the two movement cues influenced naturalness ratings, we conducted repeated measures analyses of variance (ANOVAs) with two within-subjects factors, temporal offset magnitude (0.5, 1.0 s) and offset direction (lead, lag). As shown in Figure 2.2, the results revealed a significant main effect of temporal offset direction, $F(1,94) = 8.95$, $p = 0.004$, $\eta_p^2 = 0.842$. This finding indicates that observers judged actions to be more natural when the temporal offset was consistent with the expected causal direction (lag condition) than when there was an equal amount of temporal offset in the lead condition (i.e., when the temporal offset was opposite the causal direction). As expected, we found a significant main effect of offset magnitude, $F(1,94) = 28.73$, $p < 0.001$, $\eta_p^2 = 1.0$, which indicates that people were sensitive to the general degree of temporal alignment between limb movements and body displacements when assessing the validity of observed actions, and larger offsets resulted in lower naturalness ratings. The two-way interaction between offset magnitude and temporal direction was not significant, $F(1,94) = 0.10$, $p = 0.754$. Although the rating differences in the absolute scale may appear small, it should be emphasized that (as noted previously) participants' ratings did not span the full 5-point scale. These mean ratings indicate that participants tended to provide naturalness ratings in the middle range, as long as the observed limb movements did not obviously violate biological constraints (which is consistent with results from a previous study on action recognition; Thurman & Lu, 2013).

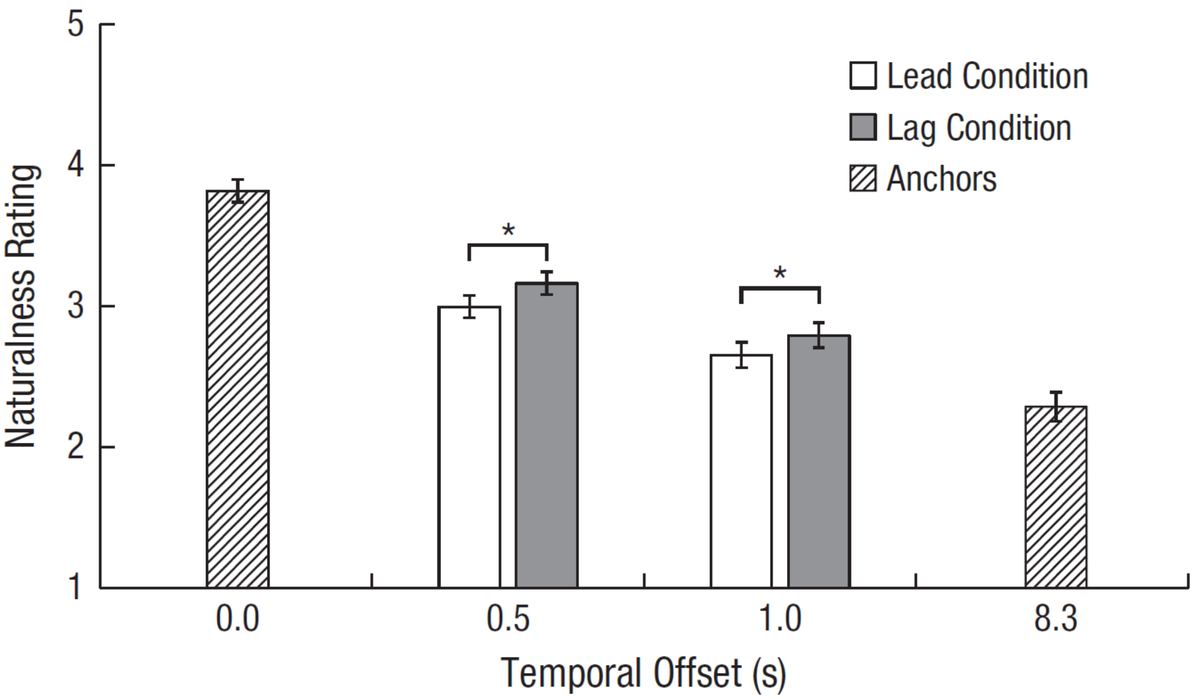

Figure 2.2: Results from Experiment 1: mean naturalness rating as a function of temporal offset between limb movements and body displacements, presented separately for the lead condition and the lag condition. Also shown are the mean naturalness ratings in the conditions with offsets of 0.0 and 8.3s, which anchored the range of ratings. Error bars represent 1 SEM. Asterisks indicate statistically significant differences between conditions (*p < .05).

2.4 Experiment 2a

In Experiment 2, we aimed to gauge the inferential aspects of causality in action perception. We drew on a design typically used in studies of causal inference to explicitly separate the causal cue and its effect. We created a reasoning task in which the two movement cues were presented by distinct visual entities in the display. This new reasoning task assessed whether people used the default causal relation to infer the binding between the two types of movements.

2.4.1 Method

Participants. Twenty UCLA undergraduate students (mean age = 20.8 years; 14 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision. The sample size was estimated on the basis of sample sizes in previous studies of causal perception in a laboratory setup (N = 14 in Scholl & Nakayama, 2002). Data collection for this experiment stopped in the week when the expected sample size was reached.

Stimuli. Four action sequences of a person walking on an uneven surface were displayed from two viewing directions with orthogonal projection. The size of the walker was a maximum of 3.2° wide by 5° high. The walker was displayed as a red stick figure (3.5 cd/m^2) on a black background (0 cd/m^2); the size of the frame was 22° by 12°. The walker appeared to walk on a treadmill by maintaining a stationary position for the average location of two hip joints at the center fixation point. A gray dot (75.9 cd/m^2; diameter = 1), tracking the position change of the body over time (as a GPS navigation system shows the location of a vehicle), depicting a GPS signal of body location in the space, moved separately according to the trajectory of body displacements. Figure 2.3 provides a schematic illustration of an example stimulus. A white fixation cross (146.5 cd/m^2) was always shown at the center of the screen. Participants used a chin rest to maintain a fixed viewing distance of 35 cm.

On each trial, action stimuli were presented for 6.67 s. The first 100 frames (i.e., 1.67 s) presented only the walker to encourage participants to maintain fixation on the walking action. The GPS dot then appeared at the center of the screen in the 101st frame and subsequently moved according to the assigned trajectory of body displacements. In the experiment, the stimuli were shown from one of four viewpoints (45, 135, 225, and 315). The order of conditions was randomized.

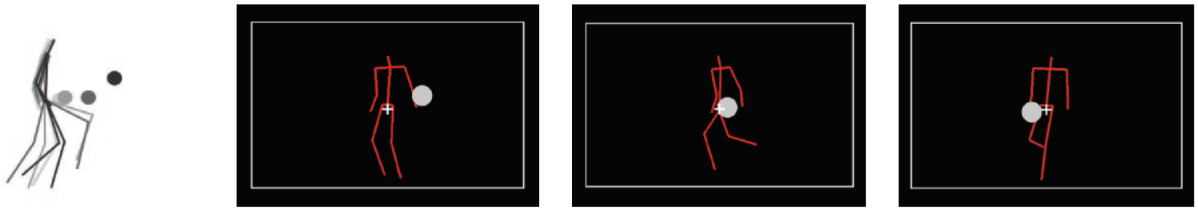

Figure 2.3: Illustrations of the stimuli in Experiments 2a and 2b. The illustration on the left shows several possible limb movements for a stick-figure walker resulting from posture changes over time. The walker remained in a stationary location (i.e., global body displacement was derived from hip movement and was then subtracted from the original walking action); the dot depicts the change in body position that results from body displacements. The sticks in the walker and the dot changed from light to dark color to denote elapsed time. Three sample frames from an experimental trial are shown on the right in to demonstrate how a dot (represented as a GPS dot in Experiment 2a and a laser spot in Experiment 2b) moved according to the assigned trajectory of body displacements with a particular temporal offset from limb movements. Videos S3 and S4 in the Supplemental Material show examples of the dynamic stimuli used in these experiments (they are also available at http://cvl.psych.ucla.edu/causal-action-2016.html).

Procedure. Participants were given a cover story: "Imagine that you work for a specialized video analysis company and are given two sources of information: (1) A posture-change video from a motion tracking system, which records a person's posture change over time and keeps the figure always at the center, and (2) A dot-motion video from a GPS system, which tracks the location of the person." Participants were then presented with a few videos demonstrating how the posture changes were separated from the change in body position over time, based on the original motion-capture video. They were informed that “It turns out that in preparing the combined videos, mistakes are sometimes made. Sometimes the posture-change video is correctly linked to the dot-motion video. However, in other cases, the posture-change and dot-motion videos were mixed up so that the two videos shown together do not match.”

Participants were asked to judge whether the posture-change video and dot-motion video matched by pressing one of the two response buttons.

Participants were given two practice blocks with feedback. Each practice block consisted of 12 trials, 6 trials of matched stimuli (i.e., temporal offset of zero) and 6 trials of unmatched stimuli with obvious temporal misalignment (i.e., body motion was 8.33 s ahead of limb movements). Matched and unmatched trials were randomly interleaved. Feedback was provided after each practice trial. For correct responses, participants were provided with a beep sound and the word “Correct” on the screen. For wrong responses, the screen displayed the word “Incorrect” at the end of the practice trial.

In the subsequent test block, 96 trials were presented to participants. The trials included eight levels of temporal offsets (0.02 s, 0.5 s, 1.0 s, and 1.5 s) between body displacements and the posture change resulting from limb movements. The experiment included 10 filler trials that came from the practice block. The 10 filler trials were randomly inserted into the experiment as attention-check trials. The entire experiment lasted for about 20 min.

2.4.2 Results

The results of Experiment 2a are shown in Figure 2.4a. A repeated measures ANOVA with two within-subjects factors (condition: lead, lag; temporal offset: 0.02 s, 0.5 s, 1.0 s, and 1.5 s) revealed a significant main effect of offset magnitude, $F(3,17) = 21.99$, $p < 0.001$, $\eta_p^2 = 1.0$, which indicates that participants were sensitive to the temporal misalignment between the two motion cues in this binding task. The interaction of offset magnitude and temporal direction of the offset was marginally significant, $F(3,17) = 3.14$, $p = 0.053$, $\eta_p^2 = 0.622$, which suggests that the influence of temporal direction between the two movement cues on the binding judgment depended on the magnitude of temporal offsets. We found that with a 1-s offset, there were a significantly higher proportion of matched responses in the lag condition ($M = 0.50$, $SD = 0.27$) than in the lead condition ($M = 0.33$, $SD = 0.27$), $F(1,19) = 9.34$, $p = 0.006$, $\eta_p^2 = 0.826$. The difference remained significant after adjusting for multiple comparisons using the Bonferroni correction procedure. A causal-asymmetry effect was thus observed within a middle range of the temporal window when the two motion cues could be interpreted as originating from a single walker but with a noticeable temporal delay between the two motion signals.

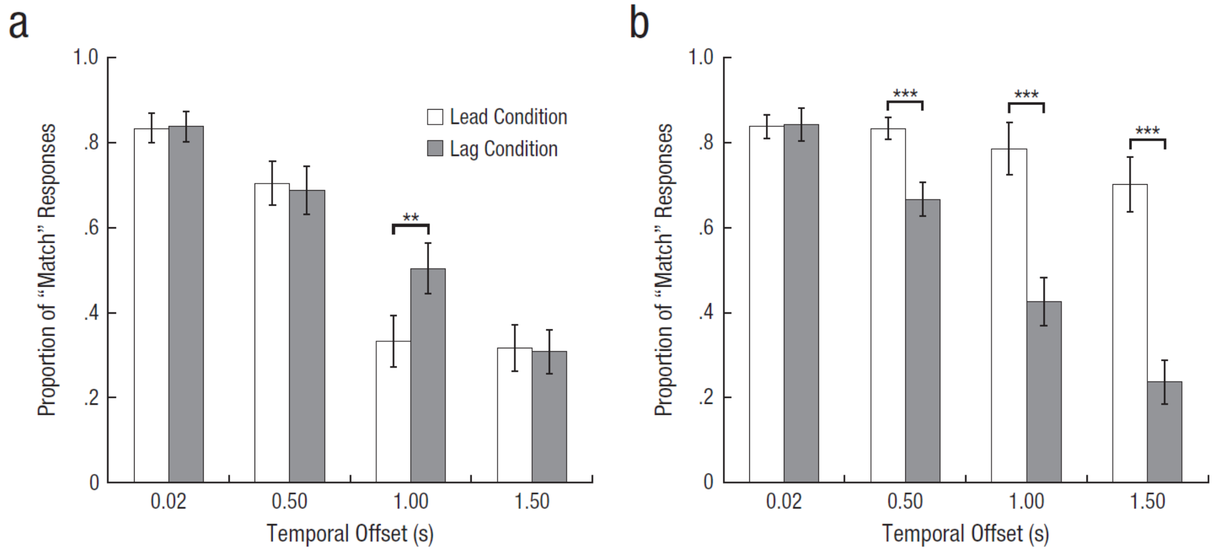

Figure 2.4: Results from (a) Experiment 2a and (b) Experiment 2b: mean proportion of "match" responses as a function of temporal offset, presented separately for the lead condition and the lag condition. In Experiment 2a, the temporal offset was between relative limb movements and body displacements. In Experiment 2b, the temporal offset was between relative limb movements

(effect) and the motion of the laser dot (cause). Error bars represent ±1 *SEM*. Asterisks indicate significant differences between conditions (***p* < .01, ****p* < .001).

A temporal-asymmetry effect was not observed for the temporal offsets in the extreme conditions. When the temporal offset was very small (e.g., 0.02 s), observers might not have detected the temporal differences between the lead and lag conditions. However, when the temporal offset was very large (e.g., 1.5 s), observers might infer that the two videos were generated from different sources; therefore, they judged the two signals to be mismatched, regardless of the temporal direction of offset. Two post hoc control experiments were conducted to test these predictions.

In one control experiment, we showed two displays side by side, each with a temporal offset of the same magnitude, but one was positive and the other negative. Eight observers were asked to judge whether the two displays were the same or different. We found that observers judged the two displays with temporal offsets of 0.02 s and –0.02 s to be the same on a high proportion (*M* = .92, *SD* = .08) of trials. In contrast, for each of the other three magnitudes of temporal offsets (0.5, 1.0, and 1.5 s), people judged them to be the same much less often (*Ms* = .22, .13, and .13, respectively). These findings support the hypothesis that people can barely detect the difference between temporal offsets of 0.02 s and –0.02 s but are sensitive to the difference in temporal direction when the magnitude of temporal offsets was 0.5 s or more.

In a second control experiment, we showed the same visual stimuli that we used in Experiment 2a. In the cover story, we introduced the participants (N = 21) to a one-person situation, in which the two sources of motion (i.e., the limb movements and the dot motions) came from the same walker, and a two-person situation, in which the each of the two sources of motion came from a different walker. The participants' task was to make a two-alternative forced choice about whether the two sources of motion came from one person or from two people. Results showed that the proportion who chose the two-person situation increased with the temporal offset (0.02-s offset: *M* = .22; 0.5-s offset: *M* = .38; 1.0-s offset: *M* = .56; and 1.5-s offset: *M* = .63). Only for the longest temporal offset (1.5 s) did the proportion of "two-person" choices significantly surpass the chance level of .50 (*M* = .63, *SD* = .18, *p* = .004). This finding indicates that when the display shows a longer temporal offset, people are likely to attribute the two motion cues to two different sources, which removes the dependence of judgments on temporal directionality.

In summary, Experiment 2a used a binding task to provide evidence that observers are sensitive to the directionality of temporal offset between limb movements and body displacements in reasoning about the relation between the two motion sources. Specifically, starting at around 1 s of temporal offset, the two motion cues were more likely to be judged as matched when the causal limb movement preceded the effect of body displacement in a direction consistent with the natural causal relation (lag condition).

2.5 Experiment 2b

If the temporal-asymmetry effect found in previous experiments resulted from observers' understanding of the inherent causal relation between the two motion sources involved in human body movements, the effect should be radically altered when the causal relation is changed. In Experiment 2b, we changed the cover story to specify that the dot represented a moving laser dot (as a cause) that was being followed by a person (as an effect). In this case, rather than the limb movements causing the dot motion, limb movements are inferred to be the effects of dot motions, so the direction of causality has been reversed from Experiment 2a. This type of manipulation of schematic understanding by a cover story has been used in many previous studies to distinguish the impact of causal interpretation from associative learning (e.g., Waldmann & Holyoak, 1992).

2.5.1 Method

Participants. Nineteen students (mean age = 20.2 years; 14 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision. None of the observers had participated in Experiment 2a. Data collection for this experiment stopped in the week when the expected sample size was reached.

Stimuli and procedure. The stimuli, task, and procedure in Experiment 2b were identical to those in Experiment 2a, except for a small wording change in the cover story, which now referred to "a dot-motion video with a moving laser-generated spot, which the person is following closely."

2.5.2 Results

In this situation, the two components of the stimuli are interpreted to represent two distinct entities, such that the moving dot (i.e., the laser dot) would now be the cause that should make the agent move in a certain way, and limb movements should be interpreted as the effect. Otherwise, stimuli were identical to those in Experiment 2a. The binding task was performed in the same manner as in Experiment 2a. Accordingly, any differences in people's patterns of judgments between Experiments 2a and 2b could be attributed only to the influence of the different causal interpretations conveyed by the cover story in each. The results of Experiment 2b are shown in Figure 2.4b.

A repeated measures ANOVA revealed a significant main effect of offset magnitude, $F(3,16) = 18.97$, $p < 0.001$, $\eta_p^2 = 1.0$, and a significant interaction of offset magnitude and temporal direction of offset, $F(3,16) = 16.09$, $p < 0.001$, $\eta_p^2 = 1.0$. When the dot motion preceded the limb movements, (which was consistent with the causal direction in the person-following-a-dot cover story), participants gave a high proportion of match responses, regardless of the magnitude of temporal offset. However, when dot motion followed limb movements (which was opposite the causal direction in the person-following-a-dot cover story), the magnitude of the temporal offset had a significant impact on human judgments. The proportion of match responses in the lead conditions was significantly greater than the proportion of match responses in the corresponding lag conditions for offsets of 0.5 s, 1.0 s, and 1.5 s (all *p*s < .001), which indicates a stronger tolerance for temporal deviations between the two motion sources when the dot motion (cause) preceded the limb movements (effect), relative to the corresponding condition in which the effect cue preceded the cause.

We conducted a mixed-model repeated measures ANOVA to examine the difference in response patterns in Experiments 2a and 2b (in which the cover story changed but the judgment task and stimuli were the same). The dependent variable in this analysis was the temporal-asymmetry effect, calculated as the difference in the proportion of match responses between the lag condition and the corresponding lead condition (i.e., the lead condition with the same offset magnitude). Results showed a significant interaction effect between the magnitude of temporal offsets and cover story, $F(3, 35) = 13.29$, $p < 0.001$, $\eta_p^2 = 1.0$. This interaction effect reflects the fact that when observers received the laser spot cover story in Experiment 2b, their judgments changed dramatically and effectively reversed their interpretation of the cause-effect relations between motion cues.

Furthermore, this temporal-asymmetry effect was maintained for a larger range of offset magnitudes (from 0.5 s to 1.5 s) for Experiment 2b than for Experiment 2a. The strength and robustness of the effect in Experiment 2b are probably due to participants' qualitative interpretation of a "following" action. Participants may perceive a causal relation between the two motion cues as long as the movement of one entity follows the same trajectory as that of another entity, without necessarily being constrained to a specific value of temporal delay between the two movements. Whereas the causal relation between limb movements of a person and displacement of their body (Experiments 1 and 2a) is theoretically more closely coupled in time, the action of an agent that is following a separate object (Experiment 2b) can be much more temporally variable, as long as the motion of the agent consistently lags behind that of the object.

In summary, Experiments 2a and 2b used identical visual stimuli but induced different causal beliefs about the relation between limb movements and a distinct moving object (the dot). The opposite pattern of judgments obtained in Experiment 2a compared with Experiment 2b suggests that when considering movements attributed to a single walker, by default observers link bodily movements of articulated limbs (causes) to motions of body locations through the environment (effects). But when considering the movements of one agent with respect to a separate distal object, observers can flexibly assign the effect role to bodily movements of the agent, interpreting them as being caused by intent to follow a moving target. The contrast between the patterns of results observed in Experiment 2a and Experiment 2b provides strong evidence that the temporal-asymmetry effect is based on the observer's attribution of the causal relation between two movement cues and does not reflect a mere association or low-level physical properties of the displays.

To assess the possibility that the temporal-asymmetry effect might be due to statistical learning of temporal regularities (i.e., that limb movements should precede body displacement in time, without necessarily assuming limb movements cause the latter motion), we performed a further post hoc control experiment that introduced a noncausal association relation between the two motion cues. In the cover story, participants were told:

Imagine that two actors aim to synchronize their body movements and they walk on two identical terrains, each with steps. You are given two sources of information: (1) a posture-change video from a motion tracking system, which records Actor 1's posture change over time, and (2) a dot-motion video from a GPS system, which tracks the location of Actor 2.

Participants were asked to judge whether the posture-change video of Actor 1 and the dot-motion video of Actor 2 matched one another, such that the two actors moved in synchrony. All other aspects of the experiment were identical to those of Experiment 2a. Twenty-two UCLA students participated in this study. In contrast to the results of Experiment 2a, none of the offset conditions revealed a difference in the proportions of "match" responses between the lag and lead conditions, including the critical 1.0-s temporal-offset condition (lead: *M* = .37, *SD* = .23; lag: *M* = .37, *SD* = .20), $t(21)<0.001$, $p = 1.0$. These findings indicate that mere temporal association was not sufficient to elicit a temporal-asymmetry effect in the absence of a direct causal link between the motion cues.

2.6 An Ideal Observer Model Based on Visual Statistics

As an additional test of the adequacy of a purely associative account of our findings, we developed an ideal observer model solely based on visual statistics of action stimuli. This model allowed us to determine whether contingent associations learned from experience could predict the temporal asymmetry effect. The observer model is a hypothetical device that makes optimal decisions given available information based on natural statistics of the visual environment (Geisler, 2011; Kersten, Mamassian, & Yuille, 2004; Lu, & Yuille, 2006). To capture the visual statistics in relevant action stimuli, we analyzed 20 walking actions from the CMU motion-capture database, in which a walker explored an indoor environment on an uneven surface. We generated a pool of point-light stimuli from 20 actions each viewed from six different directions with orthogonal projections. This stimulus set consisted of 180,000 posture frames (20*1500*6), each including the coordinates of the 13 joints involved in the action. For each posture, body displacement (v) was calculated as the position change of averaged hip joints between two neighboring frames. To quantify limb movements, we relied on posture analysis, which has been shown to be an important computational mechanism in action recognition (Theusner, de Lussanet, & Lappe, 2014; Lange & Lappe, 2006; Thurman & Lu, 2014b, 2016b). A K-mean algorithm (Jain, 2010) was used to categorize the posture frames into a finite number of key postures involved in walking on the uneven surfaces. We selected 500 posture clusters (C), as model performance reached a plateau and did not change with additional clusters. Each of the 180,000 frames in the stimulus set was assigned to the most similar posture cluster. We then computed the histogram of body velocity given each of the posture clusters, $P(v|C, matched)$, and fitted the histograms using a 2D Gaussian distribution, as illustrated in the left panel of Figure 2.5.

To simulate the performance of observers in Experiment 2a, the model needs to estimate the probability that the observed two pieces of information provided by stimuli, body velocity S_v and posture S_P, are from matched signals in actions. Assuming equal prior probability for matched and mismatched judgments, the posterior probability will be determined by the likelihood according to the empirical distribution measured above, $P(v|C, matched)$, weighted by the probability that the observed posture is sampled from each of the posture clusters, as described in Equation 2.1.

$$P(matched|S_v, S_P) \propto \sum_C P(v = S_v|C, matched)P(C|S_P) \qquad (2.1)$$

Figure 2.5 (right) shows the simulation results. Higher log-likelihood values indicate a greater probability of considering the two motion cues as “matched”. The ideal observer model consistently predicted slightly more matched responses in the condition in which the GPS dot was shifted ahead of the limb movements than when the GPS dot was shifted behind the limb movements. This result is opposite to the temporal asymmetry effect observed from human judgments recorded in Experiment 2a. This pattern of ideal observer results was obtained for a wide range of cluster numbers in the model simulation (from 20 clusters to 1000 clusters). Hence, the observer model using statistical associations derived from large numbers of observations based on past experience (but lacking any causal understanding of human walking actions) failed to account for the pattern of human judgments. This failure of the model based solely on statistical association between body displacements and postures in actions resulting from limb movements provides converging evidence that causal understanding of body movements serves as a critical constraint in action perception and understanding. This finding implies that the causal constraint does not arise from statistical associations derived by passively viewing bodily movements in the world, where we commonly observe the occurrence of the two types of motion in near synchrony. Rather, the causal constraint may be acquired through active interaction as an agent influencing the physical and social world.

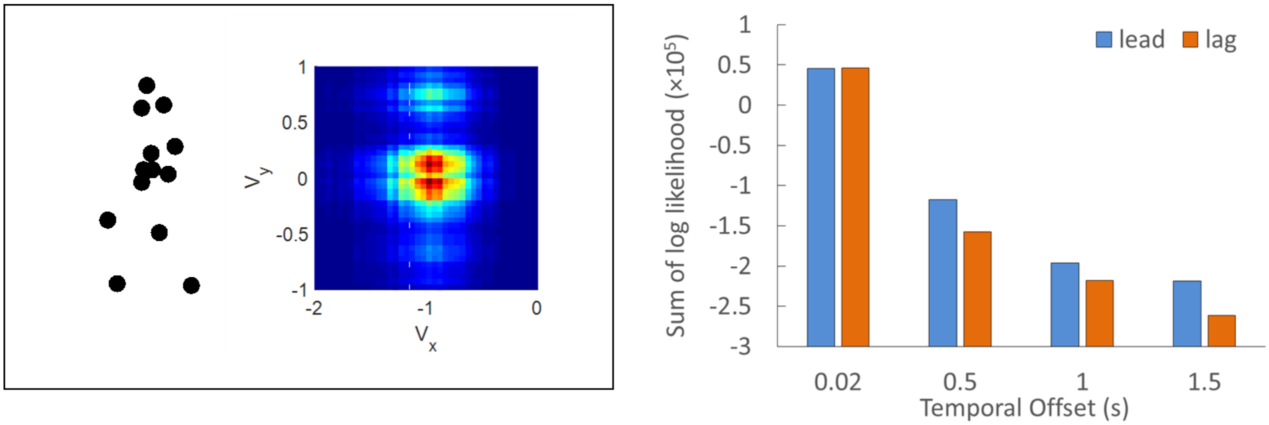

Figure 2.5. Results of an ideal observer model based on visual statistics. Left panel, distribution of body displacements (right) associated with an example key posture (left), derived from visual statistics from action observations. *Vx* and *Vy* refer to horizontal velocity and vertical velocity of body displacements (negative values indicated moving to the left on x-axis and moving down on y-axis). The distribution of body displacements suggests that the posture is more likely associated with moving towards the left, with small amount of vertical motion. Right panel, model simulation results for offsets used in Experiment 2a. These are opposite from human results, implying that the association between limb movements and body displacements learned from visual statistics is insufficient to account for human performance.

2.7 General Discussion

The basic finding in the present study is that participants were more likely to bind human actions in a causally consistent way. Although observers commonly observe the occurrence of the two types of motion in synchrony, they can tolerate deviation from this normal synchronicity.

Our results show that the degree of tolerance is constrained by the directionality of the causal relation between motions.

In fact, such a temporal-asymmetry effect has been observed (albeit not previously noted by researchers) even in the classic ball-collision paradigm introduced by Michotte (1946/1963), which elicits immediate and irresistible causal impressions that one moving ball causes a second ball to move or launch. It is well known that the causal impression is reduced after the introduction of a spatial or temporal gap. These two manipulations can be interpreted as temporal offsets with different directions in time. In the spatial-gap condition, the effect ball (the one that is launched) moves before the causal ball (the launcher) reaches the contact location (i.e., effect precedes the cause), analogous to the lead condition in the present article. In the temporal-gap condition, motion of the effect ball is delayed after the causal ball arrives at the contact location (i.e., cause precedes the effect with an abnormal temporal gap), analogous to our lag condition. We examined human data from a recent study using this paradigm (Sanborn, Mansinghka, & Griffiths 2013). Given the speeds of the moving balls used in their study, a 1-cm spatial gap corresponded to a positive 16-ms temporal offset (i.e., the effect ball moved for 16 ms before the causal ball arrived in the contact location), which yielded a causal rating of about .4 in the lead condition. In comparison, a 16-ms temporal gap yielded a causal rating of .75 in the lag condition. Thus, in conditions that equated the magnitude of temporal offset, people gave much higher causal ratings to the lag condition, presumably because it preserved the expected causal order of events.

A temporal-asymmetry effect has now been observed for a range of causal events, including object interactions, action-initiated changes in object status (e.g., a button press triggering a flashed disk; Desantis & Haggard, 2016; Rohde, Greiner, & Ernst, 2014), and body movements (as observed in the present study). We expect that similar effects would be observed for other causal events involving nonbiological stimuli, such as rotating wheels (as cause) and cars moving forward (as effect), as long as people have an understanding of the causal relation involved in the physical system. From this perspective, human body movements do not have a special status; rather, they serve as one example of causal events that yield the temporal-asymmetry effect. However, human actions may elicit a temporal-asymmetry effect with different timing properties relative to other physical systems that involve inanimate objects. For example, in the ball-collision and action-initiated-flash situations, the magnitude of temporal offset that yields an asymmetry is very short (< 200 ms), which is consistent with causal mechanisms that operate quickly and perhaps spontaneously (Scholl & Tremoulet, 2000). In contrast, the asymmetry effect found for body movements in the present study is substantially longer (offset of about 1.0 s). This larger temporal window indicates much greater tolerance for temporal misalignment between limb movements and body displacements, which may reflect the increased complexity of the causal mechanisms involved in perceiving body movements.

One possible alternative explanation of the temporal-asymmetry effect observed in the present study is that people's judgments are guided by temporal regularities learned in the environment and perhaps transferred by analogy to the similar situations described in the cover stories used in Experiment 2. For example, people are likely to have a clear expectations of temporal order for relations such as "chasing" or "following." In general, a causal relation implies a certain temporal order (cause before effect), but perhaps causality is just one special case (albeit an important one) of effects more directly attributable to temporal knowledge per se.

Although this alternative explanation might account for the findings of Experiment 2, which involved cover stories that probably conveyed temporal as well as causal knowledge, it does not offer a compelling explanation of the results of Experiment 1. Experiment 1 did not manipulate a cover story; participants simply observed point-light displays of walkers and judged their naturalness as a function of temporal offsets. In daily life, people often observe the co-occurrence of two types of motion at the same time (e.g., planting a foot, extending the legs, and moving the body forward co-occur). If learned temporal regularity, rather than causality, was the key factor, we would expect to have observed perceived naturalness peaks at zero offset, with a symmetric decline as the temporal offset was increased. Instead, we found the asymmetrical pattern predicted by a causal interpretation. Hence, the causal relation between limb movements and body displacements provides a parsimonious explanation for the findings regarding both action perception and inference. However, we note that using solitary actions of a single actor makes it difficult to isolate causation from temporal regularity (i.e., to generate conditions analogous to Michotte's temporal-gap experiment). Future studies may be able to further distinguish effects of causality versus temporal order per se by investigating patterns of motion cues involving interactions between two actors.

Actions can be interpreted as a willful expression of body movements in the environment, caused by intentional patterns of limb movement. Sensitivity to causal dynamics in body movements may play a general role in tracking perceptual animacy, which supports the ability to visually distinguish living from nonliving entities (Thurman & Lu, 2013, 2014a). In addition, as relational binding in general enhances representational power (Lu, Chen, & Holyoak, 2012), the perceived causal relation between the two types of movement cues makes it possible for people to understand why the body moves the way it does. Causal understanding of actions enables explanation of the past as well as prediction of the future (Cheng, 1997), which provides a fundamental constraint on perception and inference from human bodily movements.

CHAPTER 3:
Perception of Continuous Movements from Causal Actions

3.1 Abstract

We see the world as continuous with smooth movements of objects and people, even though visual inputs can consist of stationary frames. The perceptual construction of smooth movements depends not only on low-level spatiotemporal features but also high-level knowledge. Here, we examined the role of causality in guiding perceptual interpolation of motion in the observation of human actions. We recorded videos of natural human-object interactions. The frame rate was manipulated to yield short and long stimulus-onset-asynchrony (SOA) displays for a short clip in which a catcher prepared to receive a ball. The facing direction of the catcher was either kept intact to generate a meaningful interaction consistent with causality, or it was transformed by a mirror reflection to create a non-causal situation lacking a meaningful interaction. Across four experiments, participants were asked to judge whether the catcher's action showed smooth movements or sudden changes. Participants were more likely to judge the catcher's actions to be continuous in the causal condition than in the non-causal condition, even with long-SOA displays. Body orientation was manipulated to demonstrate the robustness of the causal interpolation effect, which held as long as the temporal contingencies in causal actions were maintained. These findings indicate that causality in human actions guides interpolation of body movements, thereby completing the history of an observed action despite gaps in the sensory information. Hence, causal knowledge not only makes us see the future but also fills in information about recent history.

3.2 Introduction

In our daily life, we are constantly incorporating new visual information to form a continuous impression of the dynamic world. However, the perceptual construction of smooth movements is not a trivial task, since visual inputs are actually discrete frames or disjointed clips separated by constant eye movements. Flipbooks, for example, exploit our susceptibility to apparent motion (Wertheimer, 1912), where our visual system induces the perception of dynamic scenes from the presentation of static images in rapid succession.

Apparent motion offers an illustrative case of the human visual system's tendency to interpolate the paths of perceptual objects over time, and to produce the perception of smooth motion across discrete samples of visual stimuli at different time points. The appearance of smooth motion is not exhaustively determined by low-level visual features, such as inter-frame spatial displacement and temporal sampling rate (Braddick, 1974; Burr, Ross & Morrone, 1986). Previous studies have demonstrated that it is only within certain ranges of displacements and stimulus-onset-asynchrony (SOA) between frames that a two-frame stimulus evokes a percept of continuous motion, and apparent motion is lost when the spatial and temporal parameters exceed those limits (Baker, Curtis, & Braddick, 1985a, 1985b; Bours, Stuur, & Lankheet, 2007; Lappin & Bell, 1976; Morgan & Ward, 1980). These behavioral findings are consistent with neurophysiological findings that neurons in early visual areas (such as V1 and MT) demonstrate similar spatial and temporal limits (Baker & Cynader, 1986; Churchland, Priebe, & Lisberger, 2005; Mikami, Newsome, & Wurtz, 1986; Newsome, Mikami, & Wurtz, 1986).

The perception of smooth motion is also influenced by high-level visual knowledge about shapes, objects and events involved in the stimuli (Sigman & Rock, 1974; Braddick, 1980; Shiffrar & Freyd, 1990; 1993; Chen & Scholl, 2016). Previous research provided compelling evidence to show that higher-level processing incorporates visual knowledge to influence interpolation of motion signals in image sequences. For example, in typical cases of apparent motion, an object presented in one frame tend to be perceived as moving with the shortest path to its location in the subsequent frame. However, Shiffrar and Freyd (1990; 1993) showed that, when two frames of human body postures are presented with long duration between frames, people perceive the body movements following a longer path that satisfies biomechanical constraints, rather than the shortest path which would imply to move cross the body. Another compelling demonstration by Chen & Scholl (2016) further illustrates that high-level knowledge can influence motion perception, even introducing illusory motion for static images. In the study, observers watched a change from a complete square shape to a truncated form with a missing piece and were asked to report whether this change was sudden or gradual. The key manipulation was whether the shape of the missing piece was generated by an "intrusion" physical interaction, e.g., as when an object pushed into a lump of clay, or whether it was imposed by cutting a piece out of clay. Results showed that observers in the intrusion condition were more likely to perceive a gradual shape change, or illusory gradual motion, even when it was actually a sudden change than did the observers in the imposed condition. This result suggests that causal knowledge probes the visual system to reconstruct motion from static images.

The typical view about causal knowledge emphasizes its application to reasoning and thinking, as knowledge about the causal structure of the world is useful for making predictions, generating explanations, and planning interventions (Cheng, 1997; Woodward, 2003). However, there is also ample evidence to show that causal knowledge can influence perception and memory. Previous evidence shows that causality between limb movements and body motions in human actions influences judgments of naturalness of actions and their inferences about human actions (Peng, Thurman, & Lu, 2017). Causal knowledge has been shown to elicit false memories of actions. Strickland and Keil (2011) found that implicit causal connections between agents and objects led to false memories of action frames that were never presented. For example, adults watched videos in which an actor kicked a ball, but the videos omitted the moment in which the actor actually contacted the ball. In a later recall task, participants falsely reported seeing the physical contact when the subsequent footage implied a causal relation between the actor's movements and the motion of the ball. Similarly, Bechlivanidis and Lagnado (2013, 2016) demonstrated that causal knowledge can induce false memories about the temporal order of events. Having a belief that event type A causes event type B made participants more likely to misremember sequences of observed events that violated those causal beliefs (i.e., when an event of type B temporally preceded an event of type A) than sequences that coincided with their causal belief.

These findings present compelling cases in which causal knowledge plays an influential role in consolidating memories about actions and events. In addition, work on causal binding has shown that causal knowledge biases the perception of time and space (Humphreys & Buehner, 2009, 2010; Buehner, 2012). For example, Buehner and Humphreys (2009) demonstrated that when one event is represented as causing another, the perceived time lapse between the two events appears shorter than when the two events are not causally related. This finding indicates that two causally related events are more likely to trigger the perception of spatiotemporal contiguity.

Here, we propose that causal knowledge of observed events not only affects visual recognition and memory, but also the visual experience of the observed events. Specifically, we examine whether high-level knowledge of the causal relations between agents and objects inherent in human activities influences the extent to which the visual system interpolates body motion. To answer this question, we test the hypothesis that higher-level visual processing incorporates prior knowledge about causal relations in human activities to fill in missing information between static frames, yielding the subjective experience of smooth motion in human actions.

We designed four experiments to examine the role of causal knowledge in guiding perceptual interpolation of motion in human actions. We recorded videos of human-object interactions in a natural environment (a thrower directing a ball to a catcher). For short clips in which the catcher prepared to receive the ball, the frame rate was manipulated to introduce short and long inter-frame durations, defined as stimulus-onset-asynchrony (SOA). The duration of short SOAs was 33.3 ms/frame; that of long SOAs was 100 ms/frame. For causal actions, the facing direction of the catcher was maintained to generate a meaningful interaction consistent with a causal interpretation. For non-causal actions, the facing direction of the catcher was reversed to disrupt any meaningful interaction and generate an action sequence inconsistent with a causal interpretation. Participants were asked to judge whether the catcher's action showed smooth body movements or sudden changes.

In Experiment 1, only the ball movement and catcher's action were shown, with the thrower occluded. We hypothesized that causal knowledge of actions influences the interpolation of discrete pieces of motion information, so that observers would be more likely to perceive smooth actions when observing causal than non-causal actions. In addition, the predicted effect is expected to be stronger for long-SOA displays in which the visual inputs are sparse, with fewer image frames. In Experiment 2 and 3, we further tested if the effect could generalize to causal human-human interactions involving both the thrower and the catcher. We hypothesized that the causal interaction between two agents would facilitate visual interpolation and induce the same effect of smooth motion perception as Experiment 1. In Experiment 4, upright videos were compared to inverted videos to test an alternative hypothesis that the effect could be simply explained by visual familiarities, such as body orientation, rather than causality.

3.2 Experiment 1

Experiment 1 was designed to assess how a causal action between an agent and a physical object influences interpolation in the perception of smooth human actions. Causal actions were generated with an agent interacting with a moving object. Non-causal actions were generated with the same agent facing away from the moving object. We hypothesized that in the causal action condition, discretized human actions would be more likely to be perceived as smooth motion sequences.

3.2.1 Methods

Participants. Fifty UCLA undergraduate students (mean age = 21.1; 40 female) participated in the experiment for course credit. All experimental procedures were approved by the Committee for Protection of Human Subjects at University of California, Los Angeles (UCLA). All participants had normal or corrected-to-normal vision.

Stimuli. Action videos were filmed with a camera in the gym with a temporal resolution of 30 frames/s. Two pairs of actors (one male pair and one female pair) were enrolled and each pair performed three throwing-catching actions (i.e. bounce pass, chest pass, and underhand throw), with each actor being the thrower once and catcher once. Seven video clips were selected as experimental stimuli. The video stimuli used in the experiments can be viewed at https://yujiapeng.com/causal-illusion-real.

In Experiment 1, only the catcher and the ball appeared in the video; the thrower was not shown. For each video, a short critical period was selected during which the catcher’s arms showed the largest rising momentum during preparation to catch the ball. Each video lasted for 567 ms. There were 10 frames before the critical period, and 1 frame after the critical period. The critical period began when the catcher's arms started to rise, and it ended right before the actor’s hands touched the ball. The duration of the critical period was 200 ms. In the long-SOA condition, only the first and the last frame of the catcher’s body movements were presented, all the middle frames were omitted. The presentation duration of the first and the last frames were lengthened to each cover half of the critical period at 100 ms per frame. In the short-SOA condition, all six frames showing body movements of the catcher were displayed, with the frame duration at 33.3 ms/frame. Note that the duration of the critical period was the same (200 ms) for both long-SOA and short-SOA displays. The movements of the ball were also the same and were kept intact in both long-SOA and short-SOA displays (Figure 3.1).

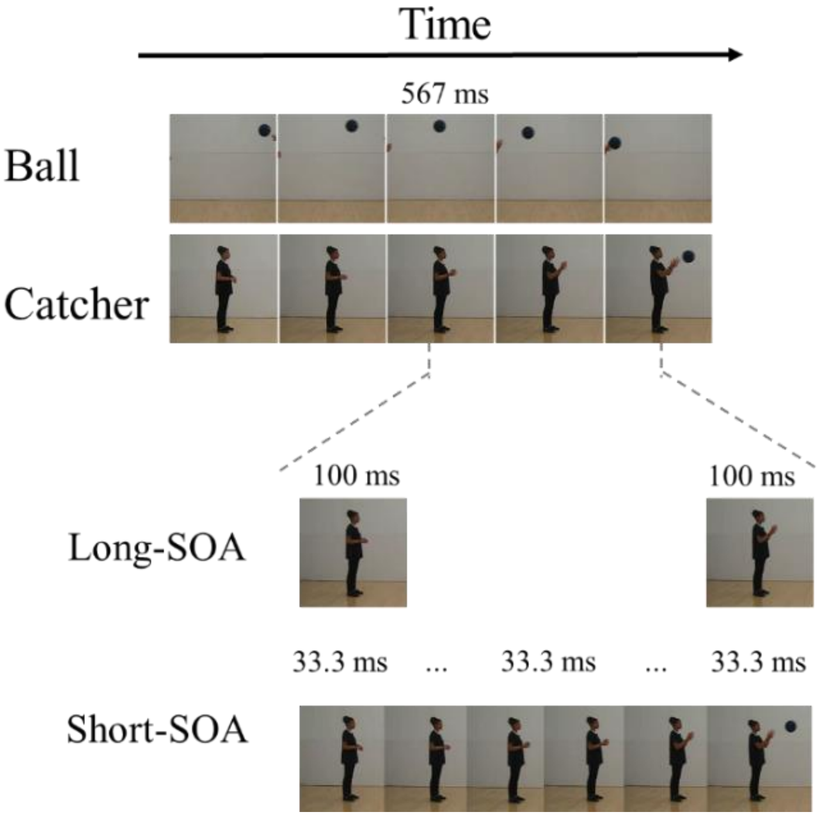

Figure 3.1. Illustrations of the critical clip in the long-SOA display with two frames (100 ms/frame) with a sudden posture change, and in the short-SOA display with six frames (33 ms/frame).

As shown in Figure 3.2, the causal condition showed the catcher facing toward the ball as the ball movement causes the catcher to move his or her body in preparation. To generate non-causal actions, image frames were processed using Matlab and Adobe Photoshop to horizontally reverse the facing direction of the catcher. The catcher was flipped horizontally to face away from the ball in the entire video, while keeping the background and the ball movement intact.

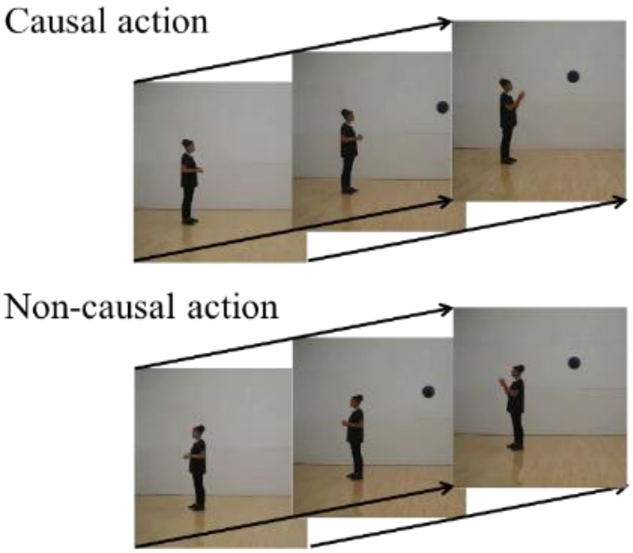

Figure 3.2: Sample frames of a causal action with the catcher facing towards the ball, and a non-causal action with the catcher facing away from the ball.

Procedure. Participants were seated 35 cm in front of a monitor with a 1024×768 resolution and 60 Hz refresh rate. All the stimuli were generated by MATLAB Psychtoolbox (Brainard, 1997). Participants were instructed, "You will view an actor playing sports (such as passing a basketball) with someone else who is occluded by a whiteboard. The task is to judge whether the catcher actor shows a smooth action or a non-smooth sudden posture change. For a smooth action, the actor smoothly moves from one posture to another. For a non-smooth action, the actor suddenly moves from one posture to another."

On each trial, a white fixation cross was presented at the center of the screen. Participants were asked to focus on the fixation cross throughout the experiment and to use their peripheral vision to see the video without making saccades. The center of the video was presented 13.7 degrees to the left or to the right of the fixation point with a height of 18 degrees. Showing the video in peripheral vision reduced the possibility that observers would track movements of the catcher without paying attention to other parts of the display. Half of the trials presented the video on the left of the fixation and the other half on the right. The catcher actor was always presented on the side relatively farther away from the fixation point. For example, if the video was presented on the right side, the ball flew from left to right and the catcher was located on the right side of the ball. After the video display, participants were asked to press one of two buttons to judge whether the video demonstrated actions with smooth body movements or sudden posture changes.

Participants were first presented with two blocks of practice trials to familiarize them with the task. In the practice blocks, participants saw “correct” on the screen plus a beep after each correct response, and they saw “incorrect” without a beep after each incorrect response. Each practice block consisted of eight trials. A separate video was used as the stimulus for the practice block; this video was not presented in the test. In the first block of practice, videos were slowed down to show the entire video with the frame rate of 66.6 ms/frame and to display the critical period for 666 ms. This manipulation was intended to allow participants to become familiar with the experimental setting and to understand the difference between smooth motion and sudden posture changes in body movements. In the second block of practice trials, videos were presented at a frame rate of 33.3 ms/frames, and the duration of the critical period was 200 ms, as in the test session.

The test session followed the practice blocks. Test trials were identical to those in the second practice block with two exceptions: participants received no feedback on test trials, and test trials employed six new videos that were not used in practice blocks. A total of five test blocks were administered, each with 24 trials (causal/non-causal x long-/short SOA x 6 actions).

In each block, the presentation order of videos was randomly shuffled. Proportions of responses in judging actions as smooth motion were recorded for each condition.

3.2.2 Results

We first examined the data in Block 1, as performance on subsequent blocks was likely to be affected by increased familiarity with the six videos used in the experiment. We conducted a 2 (SOA: short- vs. long-SOA) by 2 (causality: causal action vs. non-causal action) repeated-measures ANOVA on the proportion of responses judging the catcher's action as smooth motion. As shown in Figure 3.3, results revealed a significant main effect of causal action, $F(1,49) = 4.742$, $p = .034$. Especially for the long-SOA display, the proportion of "smooth" responses was significantly higher in the causal action condition in which the catcher faced towards the flying ball than in the non-causal action condition in which the catcher faced away from the ball ($t(49) = 2.243$, $p = .029$). The proportion of judging the long-SOA video as a smooth action increased 43.7% from the non-causal condition to the causal condition, yielding a Cohen's effect size value of $d = 0.31$. Note that the smooth motion signal was much weaker in the long-SOA display, since the stimulus included only two static postures with the largest spatial displacements. Given the visual input with large uncertainty in the long-SOA condition, the causal relation between the ball and the body movements of the catcher enhanced interpolation between the two distinct postures, resulting in more misperception of sudden posture changes as smooth body movements. These results indicate that the effect of causality on motion interpolation emerged at the very beginning of the experiment. Not surprisingly, the main effect of the SOA was significant, $F(1,49) = 124.803$, $p < .001$, as short-SOA displays provided stronger motion signals with short inter-frame spatial displacements than did long-SOA displays. The two-way interaction effect between causality and SOA was not significant, *F(*1,49) = .662, *p* = .42.

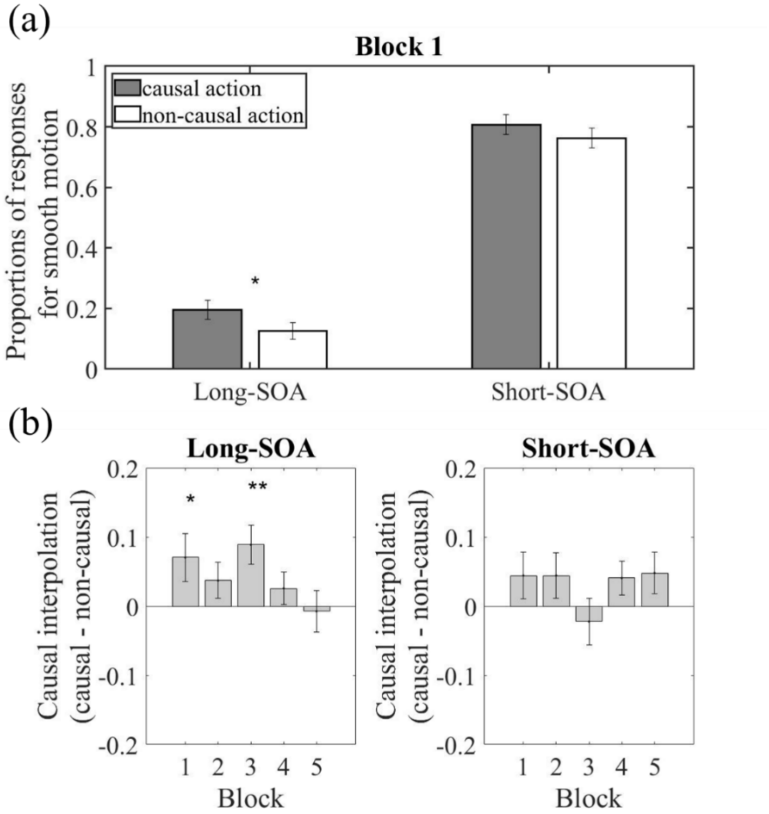

Figure 3.3: Results of Experiment 1. (a) Proportions of responses in block 1 judging the catcher's action as smooth motion. Asterisks indicate statistically significant differences between conditions (* *p* < .05, ** *p* < .01). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

To investigate whether the impact of causal actions on motion interpolation was maintained across blocks despite increased familiarity with the six videos, we conducted a three-way repeated measures ANOVA with blocks as the third factor. We found a significant main effect of causal actions (F(1,49) = 12.419, p = .001), reflecting a larger proportion of "smooth" responses in the causal condition than non-causal condition. This result suggests that the facilitation effect of causality on the perception of smooth movements was maintained, even with increased familiarity with the videos. However, the data also revealed a significant three-way interaction (F(4,196) = 2.815, p = .027), reflecting a complex relation between familiarity and the influence of causal knowledge on the perceptual task. To interpret this significant three-way interaction effect, we found that the block variable had a strong impact on responses in the long-SOA displays (F(4,196) = 4.572, p = .001), but a relatively weaker impact on short-SOA displays, for which the simple main effect of block for short-SOA was not reliable (F(4,196) = 1.722, p = .15). This pattern was likely the result of close-to-ceiling performance in perceiving smooth motion in the short-SOA displays.

3.3 Experiment 2

In Experiment 1, we found evidence that causal interactions between a catcher and the ball facilitated the perception of smooth movements. This effect was especially prominent in the long-SOA condition in which motion signals were weak. In Experiment 2, we investigated whether the effect could be generalized from human-object interactions to human-human interactivity.

3.3.1 Methods

Participants. Forty-eight new UCLA students (mean age = 20.48; 33 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli and procedure. The experiment employed the same basic videos as in Experiment 1, showing two actors pass balls. The stimuli included the body movements of the thrower and the catcher (Figure 3.4). A white occluder was presented at the center of the video to cover the movements of the ball. Depending on the actual duration of action sequences, the stimuli ranged from 633 ms to 1233 ms. There were 10 frames before the critical period, and 1 frame after the critical period. The duration of the critical period was 200 ms. In the instructions, participants were asked to respond to the movements of the catcher while paying attention to the entire video. The causal manipulation in Experiment 2 was the same as Experiment 1: the facing direction of the catcher was horizontally reversed to generate the non-causal condition. The procedure for Experiment 2 was the same as that for Experiment 1.

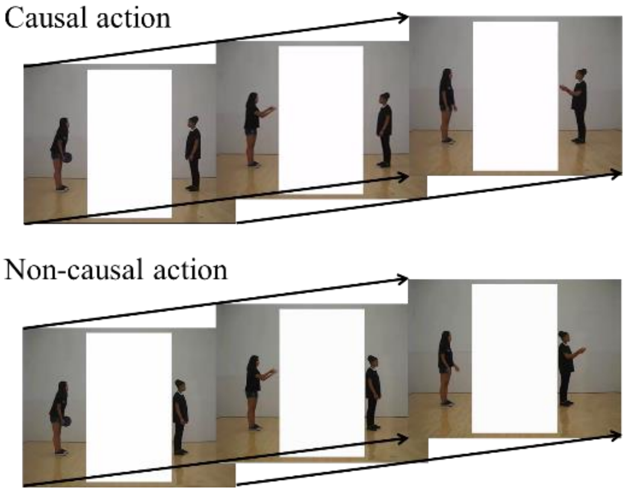

Figure 3.4: Sample frames of a causal action with the catcher facing towards the thrower, and a non-causal action with the catcher facing away from the thrower.

3.3.2 Results

As shown in Figure 3.5, the proportion of smooth responses in Block 1 again revealed a significant main effect of causality (*F*(1,47) = 9.874, *p* = .003). Despite a longer temporal delay between the two actors' actions, the causal relation between the two actors' body movements impacted the visual experience of the catcher, as perceiving the catcher's movements elicited the perception of more smooth and coherent motion. The proportion of smooth responses was significantly greater in the causal action condition compared to the non-causal action condition for the long-SOA condition (*t*(47) = 2.887, *p* = .006), but not for the short-SOA condition (*t*(47) = 1.681, *p* = .099). The proportion of judging the long-SOA video as a smooth action increased 52.5% from the non-causal condition to the causal condition, yielding a Cohen's effect size value of *d* = 0.44. No interaction effect was found, *F*(1,47) = 0.407, *p* = .527. These results extended the pattern of causal effects observed in Experiment 1.

A three-way repeated measures ANOVA with blocks as the third factor showed a significant main effect of causal actions (*F*(4,47) = 6.508, *p* = .014), with a greater proportion of "smooth" responses in the causal condition than the non-causal condition. There was also a significant main effect of block (*F*(4,188) = 5.904, *p* < .001). Neither the two-way interaction effects nor the three-way interaction effect was reliable. In summary, the converging results from the two experiments indicate that the influence of causal action on motion interpolation persisted even with increased familiarity with the videos.

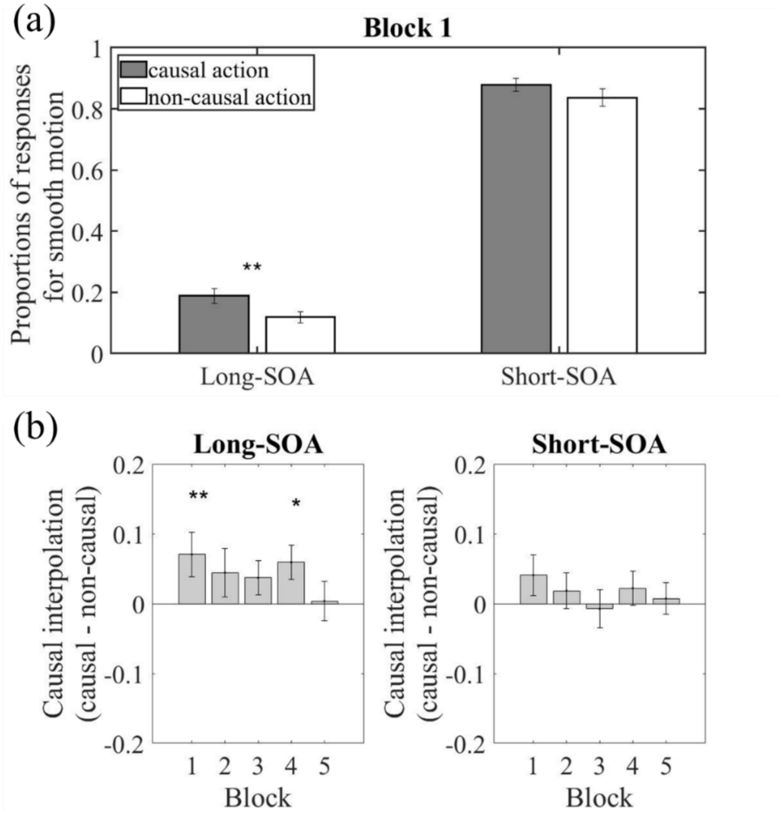

Figure 3.5: Results of Experiment 2. (a) Proportions of responses in block 1 judging the catcher's action as smooth motion (* $p < .05$, ** $p < .01$). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

3.4 Experiment 3

In Experiment 1 and 2, we found evidence that causal actions between an agent and an object, or between two agents induced stronger tendencies of perceiving smooth human body movements by interpolating information between frames. In Experiment 3, we aimed to investigate if the effect would be altered in a more complex interaction scenario when the actions of the thrower and the movements of the ball were both included. This would provide a longer causal chain, with the ball being both an effect of the thrower's action, as well as a cause of the catcher's movements. We aimed to examine if the complete causal chain would elicit a stronger effect of causal interpolation.

3.4.1 Methods

Participants. 49 students (mean age = 20.31; 39 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision. None of the participants had participated in previous experiments.

Stimuli and procedure. Same stimuli were used as Experiment 2 except that the movements of the ball were not occluded in Experiment 3. In the instructions, participants were asked to respond to the movements of the catcher while paying attention to the entire video (Figure 3.6). The procedure for Experiment 3 was the same as that for Experiment 2.

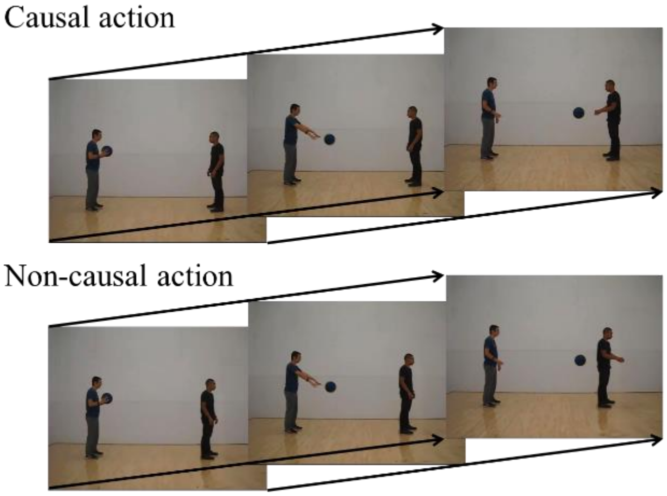

Figure 3.6: An illustration showing sample frames of an upright and an inverted action in Experiment 3.

3.4.2 Results

A 2 (SOA) by 2 (causality) repeated-measures ANOVA was conducted with the dependent variable being the proportion of judging the perceived action as a smooth action for the data in Block 1. The general findings from the first two experiments were replicated in Experiment 3. Results (Figure 3.7) showed a significant main effect of causality, $F(1,48) = 6.345$, $p = .015$, suggesting that the causal condition yielded a significantly higher proportion of judging actions as smooth actions compared to the non-causal condition. The main effect of SOA was again highly significant, $F(1,48) = 227.289$, $p < .001$. The interaction effect was not significant, $F(1,51) = .155$, $p = .696$. However, Experiment 3 showed that the proportion of rating a video as smooth was not significantly greater in the causal condition compared to the non-causal condition for the long-SOA videos ($t(48) = 1.336$, $p = .188$), but was significant for the short-SOA videos ($t(48) = 2.085$, $p = .042$). This result pattern was different from findings in Experiment 1 and 2 in which the differences showed in the long-SOA condition, not in the short-SOA condition. In Experiment 3 when the chain of causal actions was presented, as the thrower's action caused the ball to move, and then the flying ball caused the catcher to move his/her body, the significant influence of causal relations was revealed in the short-SOA condition. This was probably due to a difference in the task difficulty between experiments. In Experiment 3, action sequences generated from the entire causal chain revealed more visual information than previous two experiments, yielding increased task difficulty. We found that Experiment 3 showed significantly longer response time in judgments averaging across all conditions ($M = 914.38$ ms, $SD = 470.02$) compared to Experiment 2 ($M = 678.14$ ms, $SD = 222.78$) ($t(95) = -3.173$, $p = .002$). The increased difficulty of Experiment 3's task may have been great enough to elicit the causal effect on motion interpolation at an intermediate performance level with a short SOA display, but it may have been too great to elicit the effect at the long-SOA condition, with performance approaching a floor level.

A three-way repeated measures ANOVA with blocks as the third factor showed a significant main effect of causal actions ($F(1,48) = 20.869$, $p < .001$), with a greater proportion of "smooth" responses in the causal condition than the non-causal condition. The main effect of block was not significant ($F(4,45) = .657$, $p = .625$). Neither the two-way interactions nor the three-way interaction was reliable. In summary, the converging results from the three experiments indicate that the influence of causal action on motion interpolation persisted even with increased familiarity with the videos.

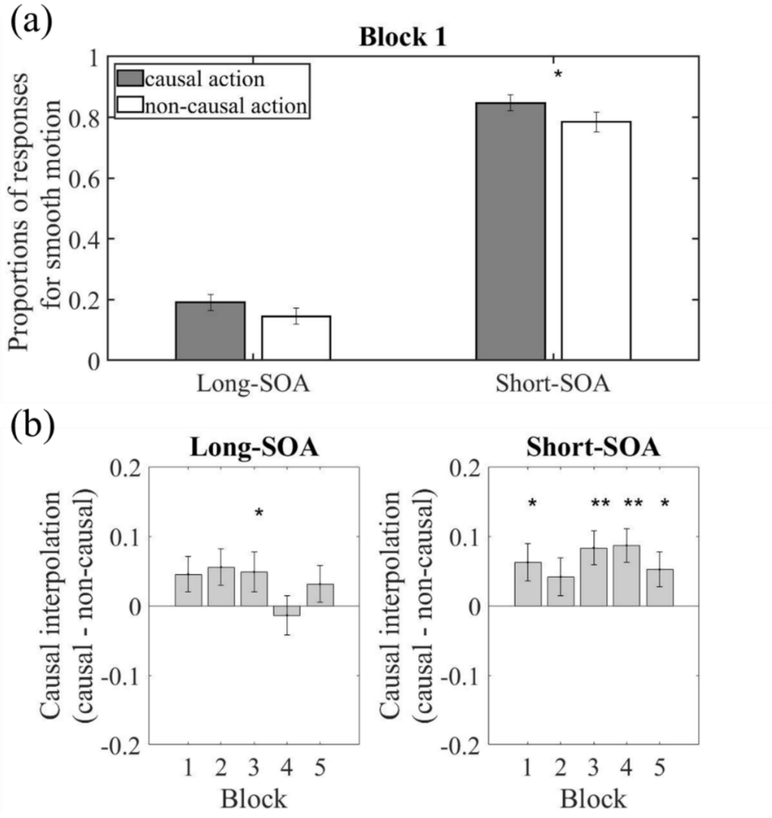

Figure 3.7: Results of Experiment 3. (a) Proportions of videos in block 1 judged as smooth actions (* $p < .05$, ** $p < .01$). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

3.5 Experiment 4

Experiment 4 aimed to investigate whether the influence of causal actions on motion interpolation depends on other visual cues.

Body orientation is a well-known cue for action recognition (Pavlova & Sokolov, 2000), as observers show worse recognition performance when actions are presented upside-down. If the interpolation effect revealed in the previous three experiments was induced by high-level causal knowledge, then inverting the video would not yield a significant difference between upright versus upside-down actions, since both cases preserve the temporal contingency and the causal relation between humans and objects. However, if the effect was merely due to visual familiarity to action interactivity, we would expect that familiar actions with upright body orientation may elicit a similar effect in comparison with unfamiliar actions with inverted body orientation.

3.5.1 Methods

Participants. Fifty-two new UCLA undergraduate students (mean age = 20.0; 43 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli and procedure. Experiment 4 used the same stimuli as the causal condition in Experiment 1. On half of the trials, the stimuli used inverted videos, and the other half used intact videos (Figure 3.8). The task and procedure of Experiment 4 were otherwise the same as in Experiment 1.

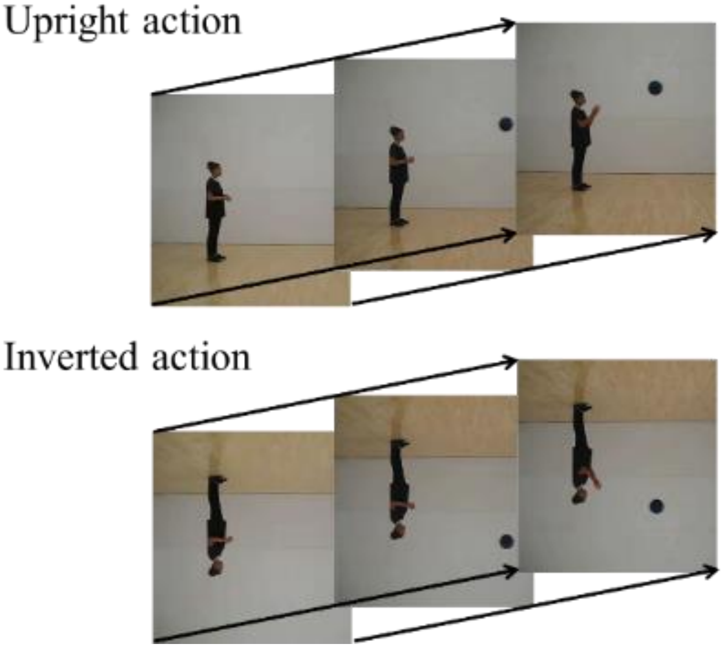

Figure 3.8: An illustration showing sample frames of an upright and an inverted action in Experiment 4.

3.5.2 Results

We first conducted a 2 (SOA: short- vs. long-SOA) by 2 (orientation: upright vs. inverted) repeated-measures ANOVA on the proportion of responses in Block 1 judging the catcher's action to be smooth motion. As shown in Figure 3.9, the main effect of orientation was not significant (*F(*1,51) = 2.509, *p* = .119). The interaction between body orientation and SOA was not significant either (*F(*1,51) = 1.525, *p* = .222). The lack of difference between upright and inverted actions suggest that as long as the causal relation and the temporal contingency is maintained in observed activities, body orientation does not affect the misperception of seeing smooth movements, even when the motion signals were weak (in the long-SOA displays).

To investigate whether the impact of body orientation on motion interpolation changed across blocks with increased familiarity with the six videos, we further conducted a three-way repeated measures ANOVA with blocks as the third factor. This analysis revealed a significant main effect of orientation (*F*(1,51) = 5.554, *p* = .022). This main effect was largely driven by a significant difference between the upright and inverted conditions in later blocks. For example, in the final block (Block 5), a greater proportion of "smooth" responses was made in the upright conditions than the inverted conditions for the long-SOA condition (*t*(51) = 2.139, *p* = .037). This pattern suggests that the impact of body orientation on visual analysis of actions increased with familiarity of the experimental videos.

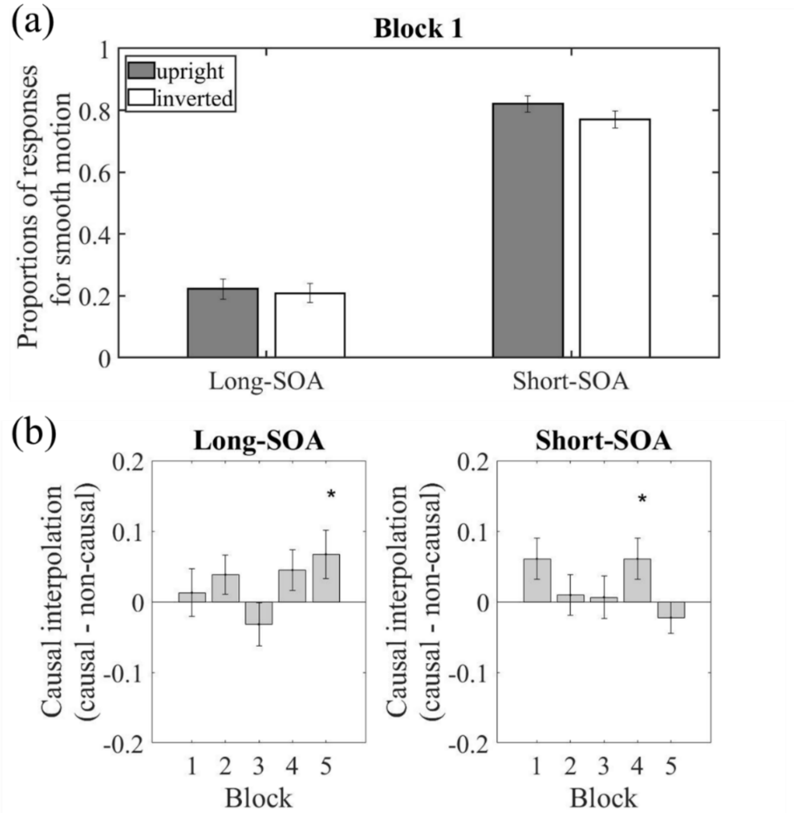

Figure 3.9: Results of Experiment 4. (a) Proportions of responses in block 1 judging the catcher's action as smooth motion (* *p* < .05, ** *p* < .01). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

3.6 General Discussion

The present paper reported converging evidence that causal relations between an agent and a physical object, or between agents increased the likelihood that people would perceive smooth actions even when the stimuli showed a sudden change between two frames. This result suggests that causality acts as a temporal “glue” to fill in observers’ visual experience by interpolating discrete image frames to produce the perception of smooth, continuous motion. The reported pattern of results can be broadly explained by invoking higher-level visual processing within Braddick’s (1980) two-process theory of apparent motion. Here, prior knowledge of causal relations involved in human actions is incorporated in higher-level visual processing, so that the recognition of events as causally connected facilitates the production of smooth motion from discrete visual inputs. The top-down influence of causality may be stronger in situations where uncertainty about the visual input is high, as we found the relatively large effect size in Block 1 of the experiment. We expect that this effect would be prominent when dynamic stimuli are presented in peripheral vision or embedded in noise. The effect may be weakened after repetitive exposures to the stimuli, as perceptual learning may enhance performance for visual tasks.

The main findings in the present paper are consistent with previous evidence that a causal understanding of observed human actions helps to fill in important visual information left out from a sequence of events and to form a continuous perception (Strickland & Keil, 2011). The current result is also consistent with previous findings of the impact of causality on the perception of shapes. The representation of an object’s implicit causal history has been shown to induce a transformational apparent motion (Tse, Cavanagh, & Nakayama, 1998) of simple objects (Chen & Scholl, 2016), akin to the “causal filling in” effect reported by Strickland and

Keil (2011). The inference of implicit causal history of objects not only changes the motion perception but also essentially has an impact on the visual shape representation (Spröte, Schmidt, & Fleming, 2016).

The impact of causality on continuous movements is potentially related to the *temporal binding* and *spatial binding. Temporal binding* is a well-documented phenomenon where the time between two events appears shorter as a function of some relation between those two events (Buehner & Humphreys, 2009; Engbert & Wohlschläger, 2007; Engbert, Wohlschläger, Thomas, & Haggard, 2007; Humphreys & Buehner, 2009, 2010; Moore & Haggard, 2007; Wohlschläger, Haggard, Gesierich, & Prinz, 2003). Haggard, Clark, and Kalogeras (2002) were the first to demonstrate this phenomenon, and they interpreted this phenomenon by appealing to a coupling between the visual system and the motor system where the temporal binding of actions and effects heightens their association in order to facilitate action-outcome learning (Haggard, Aschersleben, Gerke, & Prinz, 2002). Later, Buehner and Humphreys demonstrated that the crucial relation between the two events is causal: When one event is represented as causing another event, the time between the two events appears shorter than when the two events are not causally related. Similarly, *spatial binding* is where two objects appear closer in space when they are causally linked than when they are not (Buehner & Humphreys, 2010). Buehner and Humphreys (2009, 2010) explain both of these phenomena by invoking their theory of Bayesian ambiguity reduction. Appealing to Bayes Theorem, Buehner and Humphreys reason that two causally related events are more likely to instantiate spatiotemporal contiguity. They argue that the perceptual system uses prior knowledge of causal relations to help resolve ambiguities faced with taking noisy perceptual input to produce the subjective experience of visual motion. As result, event kinds in causal relationships are more likely to appear bound in time and space.

Causal knowledge about human body movements may not only help to connect discrete events in the perceptual process, but it may also facilitate the process of making inferences and predictions about actions. A causal framework may help the visual system to infer the past. For example, human observers get a vivid feeling of seeing the immediate past of objects or human postures presented in static frames (Kourtzi, 2004). This phenomenon suggests that causal knowledge aids the visual system in inferring and reconstructing the causal history of objects and human actions. In addition, causality in human actions may also help the visual system to predict the future. Su and Lu (2017) used skeletal biological motion displays and found a flash-lag effect, such that when a briefly-flashed dot was presented physically in perfect alignment with a continuously-moving limb, the flashed dot was perceived to lag behind the position of the moving joint. This finding suggests that the representation of human actions is anticipatory, due to a potential top-down action prediction mechanism. It has also been found that infants as young as five months are able to gaze toward the future direction implied by the static posture of a runner (Shirai & Imura, 2014, 2016), suggesting the early emergence in infancy of an ability to predict dynamic human actions from still pictures. A "causal filling in" mechanism could have benefitted from evolutionary selection pressure by aiding the continuous perception of animal motions despite occlusion by trees or other obstacles.

Recognition of causal connections in human interactions also helps the process of visual reconstruction in a top-down manner. Previous research has shown that the presence of one agent that is demonstrating communicative actions increased the likelihood of detecting a second agent's movements embedded in noise, or the "second-agent effect" (Neri, Luu, & Levi, 2006); Manera, Del Giudice, Bara, Verfaillie, & Becchio, 2011). The improvement can be explained in a framework of causal actions, where the perception of others' action is constructed not only from the visual input, but also from the intrinsic predictive activities. The presence of one agent in a causal interaction impacts the prior expectation of seeing the second agent. When the expectation derived from predictive coding is strong enough, it elicits the illusory perception of a second agent even without the valid bottom-up visual input, called “Bayesian ghost” by Manera et al., (2011).

The current study provides evidence of the important role played by causal knowledge in the perception of smooth motion. Causal relations involving human actions, and their interactions with objects and other agents, have a strong influence on motion perception for body movements. Recognition of causal relations involved in actions facilitates visual interpolation of discrete dynamic events to provide a continuous perception of human-involved activities, where the influence of causality in higher-level visual processing interacts with low-level visual processing in action perception. Hence, causal knowledge serves as a basis for the visual system’s anticipation of future events, as well as its retrospection of recent past events.

CHAPTER 4

A motion consistency constraint on infant perception of body movements

4.1 Abstract

The displacement of our body is constrained by how we move our limbs. This motion consistency between displacements of body position and limb movements plays an important role in human action perception. To examine the development of the motion consistency constraint in the perception of body movements, we recorded looking time to action stimuli in infants between 9 and 18 months of age. We compared looking time to normal actions with matched body translation to "moonwalk" actions with reversed body translation. On the first trial, infants showed a novelty preference for moonwalk actions. Longer looking times to moonwalk actions correlated strongly with walking-related motor development. These findings suggest that by 18 months, infants are sensitive to the motion consistency constraint that governs human actions, and that 9 to 18 months may be a period of substantial development during which motor functions and action perception emerge in parallel.

4.2 Introduction

In their life activities, humans perform complex movements with immense variations; nonetheless, fundamental constraints govern how people move their bodies. Physical constraints (e.g., gravity, friction) and biological constraints (e.g., limb structure) collectively determine the plausibility of observed body movements (Thurman & Lu, 2013; Shiffrar & Freyd, 1990). Functional and social goals further modulate the efficacy of body motion for interacting with the physical and social world (Pavlova, 2012). These sets of constraints work collectively to guide human body movements. Consider walking as an example. Due to biological and functional constraints, swinging the arms in an opposing direction with respect to the legs helps to reduce the angular momentum of the body motion, thereby keeping the body balanced. Due to physical constraints, only appropriate leg movements that make contact with the ground can propel the human body to walk in particular directions. Thus we expect to observe consistency between two motion signals: relative movements of limbs within an object-centered reference frame, and body displacement (i.e., translation of the body) within an environmental reference frame. Finally, due to a functional/social constraint, interaction with another person or object typically requires that one’s body be moved to a nearby location.

Numerous studies have documented that the mature visual system of adults is sensitive to the violation of fundamental constraints governing human body movements. For example, when actions are presented in an inverted orientation, physical constraints based on gravity are violated. As a result, people show poorer recognition performance for inverted actions (Sumi, 1984; Pavlova & Sokolov, 2000). When the body structure constraint is infringed by spatial scrambling, so that each moving joint is placed at a random location but maintains the same motion trajectory, recognition performance decreases (Troje & Westhoff, 2006), and perception of animacy and of social interactivity are significantly weakened (Chang & Troje, 2008; Thurman & Lu, 2014a).

Given the importance of these constraints for the recognition (as well as performance) of human actions, it is essential to understand the developmental trajectory of how these constraints are incorporated into action processing. Research suggests that some constraints on visual processing of actions may be present at birth. Two-day-old newborns exhibit a visual preference for an upright hen depicted by a point-light display over a hen shown in an inverted orientation (Simion, Regolin, & Bulf, 2008). Newborns also show greater sensitivity to biological than non-biological motion (Bidet-Ildei, Kitromilides, Orliaguet, Pavlova, & Gentaz, 2013). However, the impact of other constraints on action processing may develop after experience. Two-day-old infants do not exhibit spontaneous preference for a point-light display showing biological movement as compared to a spatially scrambled display (Bardi, Regolin, & Simion, 2011). But by 3-5 months, infants are able to discriminate between a point-light walker and a spatially scrambled display (Bertenthal, Proffitt, & Kramer, 1987), suggesting a sensitivity to the constraint of body structure. Hence, the impact of the body structure constraint on action recognition is acquired and strengthened over the course of development.

The present paper aims to examine the development of the motion consistency constraint in action perception. This constraint is based on the fact that people move their limbs in an appropriate way so as to cause the body to be displaced in the environment. When an action violates the motion consistency constraint, perception and inference of actions can be disrupted. One classic example in which this causal link is violated is the “moonwalk” dance move popularized by the dancer Michael Jackson. While the limb movements of the dancer appear to simulate walking forward, the whole body glides seamlessly *backward*, creating a dramatic conflict with the expected relations between limb movements and body displacement.

Adults show sensitivity to motion consistency between displacements of body position and limb movements (Masselink & Lappe, 2015; Peng, Thurman & Lu, 2017). The sensitivity to body translational displacement may constitute a critical first step in the development of visual perception of human actions. Three-day-old infants demonstrate a preference for a point-light walker showing translational displacement of the body over a point-light actor walking in place (e.g., on a treadmill; Bidet-Ildei. et. al., 2013). This finding highlights the apparent role that body translational displacement may play in an innate visual competency for action perception. However, it remains unclear whether infants show this visual preference as long as some sort of body translation occurs, or whether they are sensitive to consistency between body displacement and appropriate limb movements—that is, the *motion consistency constraint*—and if so, whether such sensitivity interacts with the infants' own motor development.

The current study addresses the question of whether infants discriminate human actions that satisfy versus violate the motion consistency constraint. To keep all relevant factors matched in our stimuli, both versions of actions included the same kinematic cues for the same actor and showed body translation. However, the normal action condition showed the correct binding between limb movements and body translational displacement, whereas the "moonwalk" actions reversed the horizontal motion of body translations to violate the expected walking direction. In addition, we examined whether sensitivity to the motion consistency constraint in action perception correlates with the development of gross motor functions. We included actions that are within (e.g. walking) and beyond (e.g. long jumping) the range of gross motor capabilities of infants. Thus, the experiment investigated whether, by 18 months of age, infants show a visual sensitivity for human actions with matched body displacement and whether such sensitivity correlates with the development of motor function.

4.3 Method

Participants. A sample of 34 healthy infants (20 male) between 8.8 and 18.4 months (M_{age} = 12.80 months, SD_{age} = 3.06 months) completed the study. Parents and caregivers were contacted by telephone from lists of birth records provided by Los Angeles County. Parents were given a small gift (a toy or a t-shirt) for their participation.

Stimuli and Apparatus. Action stimuli were generated from the Carnegie Mellon University Motion Capture Database (http://mocap.cs.cmu.edu) and processed using the Biological Motion Toolbox (van Boxtel & Lu, 2013). We selected four actions with whole-body movements. Two of these actions could not be performed by 18-month-old infants (non-performable actions) and two actions are typically within the capacity of the gross motor function of 18-month-olds (performable actions). Non-performable actions were “long jump” and “side-walk.” Performable actions were “walk and turn” and “walk and turn-in-place.” Each action lasted 16 seconds. Skeletal displays were generated by connecting 13 main joints of the body, including the head, shoulders, elbows, wrists, hips, knees and ankles. The size of image frames was 17.1˚ by 21.3˚ visual angles. The lower half of image frames included a checkerboard floor (see Figure 4.1). The height of skeletal actors was normalized for all actions as 9.5˚ visual angle. The vertical locations of the hip points were placed at the vertical center in the first frame of the video. The average horizontal locations of the hip points in each video were kept at the horizontal center. Body displacements were computed as the change in the average position of the two hip joints in time, and limb movements were defined as the residual motion after subtracting body displacements on a frame-by-frame basis (Peng et al., 2017). For each of the four actions, a moonwalk action was generated by reversing the horizontal direction of body displacements (see an illustration in Figure 4.1, left panel). Speeds of actors were controlled to be the same across normal action and moonwalk conditions. The stimuli used in the experiments can be viewed at *https://yujiapeng.com/infant-moonwalk*.

To measure the gross motor functions of infants, a questionnaire was developed by selecting a subset of 10 questions from the Early Motor Questionnaire (EMQ) (Libertus & Landa, 2013). The 10 questions measure infants' gross motor abilities in sitting, standing, walking, running, and jumping (see Appendix A). The measurements are based on reports from caregivers on a scale from -2 to 2, with -2 referring to "Sure that child does NOT show behavior" and 2 referring to "Sure that child shows this behavior and remember a particular instance."

Procedure. Before the experiment, a caregiver provided informed consent and completed the gross motor questionnaire. During testing, each infant sat on a caregiver's lap approximately 55 cm from a 61.5-cm monitor. Parents were instructed to hold their infant on their lap and allow their child to look freely during the session. They were also asked not to talk to their infant, point to the screen, or otherwise influence their infant's looking pattern. Prior to testing, each infant's gaze was calibrated using the standard calibration routine for the eye tracker.

On each trial, one normal action with matched body translation and one moonwalk action with reversed translation were shown side-by-side for 16 seconds (see Figure 4.1, right panel). The facing directions of the actors in the two videos were the same. The experiment included 16 trials. Each action pair (normal vs. moonwalk) repeated four times, including two trials with leftward facing direction, and two trials with rightward facing direction. The action display order was randomized into 8-trial blocks, so that all four action pairs were presented twice in the first half and twice in the second half. Half of the trials showed the normal action with matched translation on the left side and the other half showed the moonwalk action on the left side. An attention-getter was shown at the beginning of each trial to orient the infant's gaze to the center of the screen. The experimenter controlled the progression of between-trial breaks, moving on to the next trial only once the child had fixated on the attention-getter. Looking time and eye movements were recorded by an EyeLink 1000 eye-tracker (SR Research, Ltd).

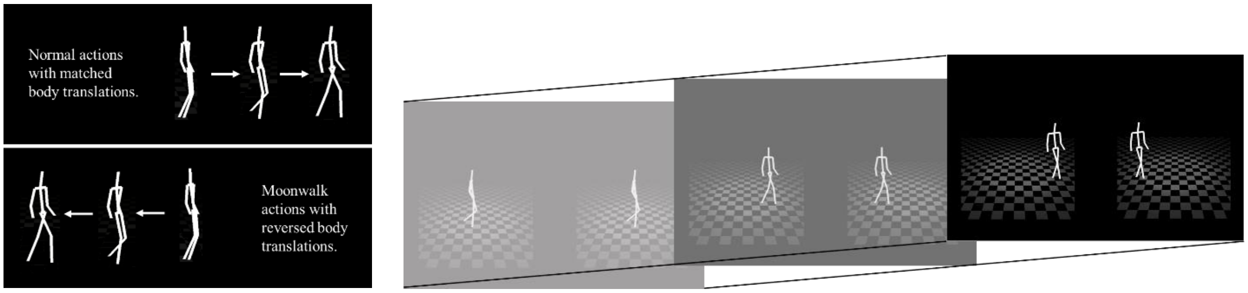

Figure 4.1. Stimulus illustration. (Left) An illustration of a normal action with matched body translations and a moonwalk action with reversed body translations. Limb movements were the same in two conditions except that in the moonwalk condition, body translations were reversed in the horizontal direction. (Right) Sample screen frames in the experiment. Lighter to darker represents frames from earlier in time to later in time. In each frame, a normal action with matched body translations and a moonwalk action with reversed body translations were presented side-by-side. Human actions were displayed in the skeletal form on a checkerboard background representing the floor.

4.3 Results

Total looking time for each action was recorded. We also examined looking proportions, calculated as the amount of time spent attending to one action divided by overall looking time to the screen for that trial. Looking time and proportion data yielded similar results; here we report findings using looking time as the dependent measure.

The mean duration of the experiment was 4.8 minutes to complete all 16 trials. The infants' looking time to the screen revealed a strong dependency on trial number. A repeated-measures ANOVA for total looking time (i.e., the total looking time with eye movement trajectories on the screen in each trial) showed that the average looking time decreased significantly across trials, reflected in a significant main effect of trial number ($F_{15,\ 19} = 7.185$, $p < .001$, $\eta_p^2 = 1.0$). Similarly, the number of infants who looked at the screen for more than half of the stimulus duration (8 s out of 16 s) on a trial dropped quickly after the first half of the experiment (Figure 4.2, left), suggesting a fatigue effect or lack of attention in the later experimental trials. In addition, exposure to skeleton displays appeared to lead to fast learning, which interacted with the visual preference to familiar versus novel stimuli. We examined the first three trials, in which most infants (>80%) looked at the display screen more than half of the trial duration. A repeated-measures ANOVA with trial number (1 to 3) and action type (i.e., normal vs. moonwalk actions) as within-subject variables yielded a significant interaction effect ($F_{2,\ 31} = 3.711$, $p = .036$, $\eta_p^2 = 0.627$) (see Figure 4.2, right). Infants' looking time showed a novelty preference (longer time to moonwalk) on the first trial, then started to show a preference to familiar displays (longer looking time to normal actions). This pattern suggests that sensitivity to the motion consistency constraint interacted with learning experience, producing a fluctuation of the preference to novel and familiar displays. In order to focus solely on the impact of the motion consistency constraint in actions, prior to any opportunity for learning or habituation, in the following analyses we focus on looking time in the first trial of the experiment.

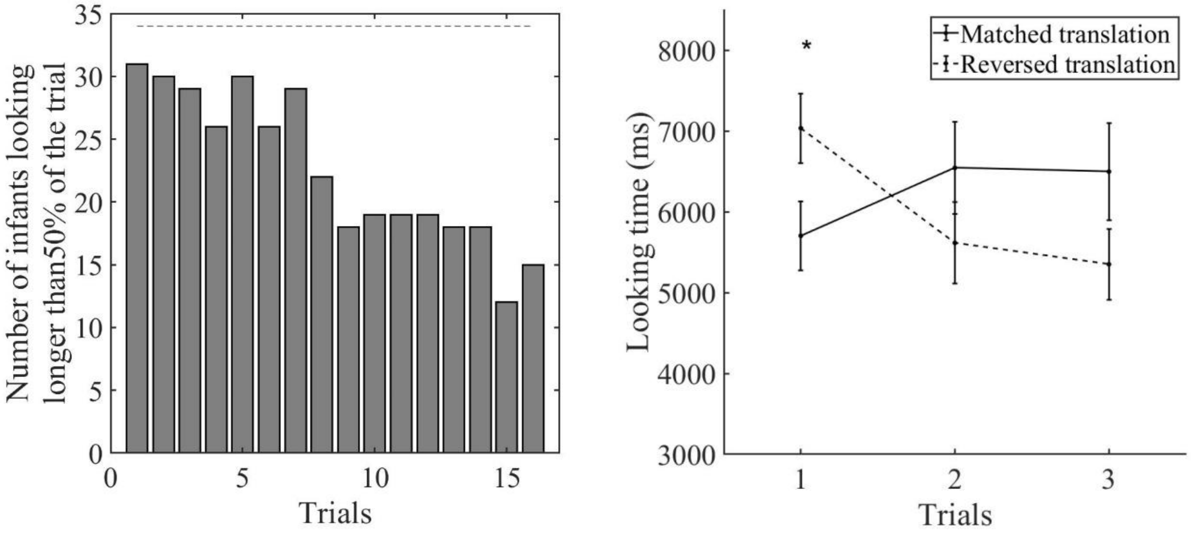

Figure 4.2. Looking time as a function of trial number. (Left)The number of infants looking longer than 50% of stimulus duration on each trial. (Right) Looking time on the first 3 trials. The preference for normal actions versus novel actions (moonwalk) varied across trials.

Effect of motion consistency and performability on first-trial looking time. A mixed ANOVA was applied to looking time on the first trial with action type (actions with matched body translation vs. actions with reversed body translation) as a within-subject variable and performability as a between-subject variable (Figure. 4.3). On the first trial, fifteen infants viewed one of the two walking videos as performable actions, and the other 19 infants viewed either long-jump or side-walk as non-performable actions. On the first trial, infants looked considerably longer at moonwalk actions (*M* = 7036 ms, *SD* = 2505) than at normal actions with matched translation (*M* = 5705 ms, *SD* = 2496), as revealed by a significant main effect of action type ($F_{1,\,32}$ = 4.298, *p* = .046, η_p^2 = .520). The longer looking time to novel stimuli of moonwalk actions indicated that infants were sensitive to violation of the motion consistency constraint.

The main effect of performability was not significant ($F_{1,\,32}$ = 0.080, p = .779). The interaction between action type and performability was not significant but marginal ($F_{1,\,32}$ = 2.872, p = .10, η_p^2 = .376). For performable actions, infants exhibited a preference for looking longer at moonwalk actions with reversed translation *(M* = 7635 ms, *SD* = 2822) than normal actions with matched translation *(M* = 4958 ms, *SD* = 2462), *t*(14) = 2.30, *p* = .037. However, for non-performable actions, infants did not show a differential preference for either normal or moonwalk actions, *t*(14) = 0.31, *p* = .76.

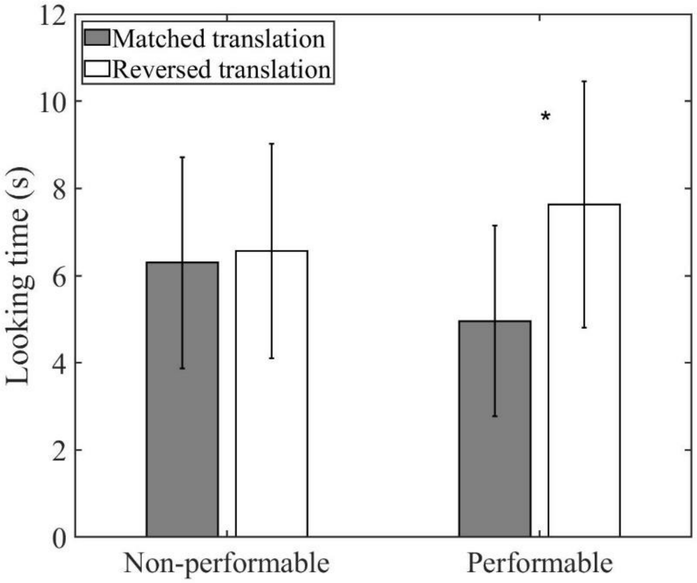

Figure 4.3. Pattern of looking time on the first trial. Infants looked longer at moonwalk actions for performable actions but not for non-performable actions. Error bars indicate standard error of looking time.

The relation between gross motor functions and looking times. A composite score of gross motor skills was calculated for each infant by summing the standardized scores of the 10 questions on the motor questionnaire. In our sample, boys were younger than girls, $t(32) = 2.12$, $p = .041$, (boys, $M_{\text{age}} = 11.91$ months, $SD_{\text{age}} = 2.63$ months; girls, $M_{\text{age}} = 14.06$ months, $SD_{\text{age}} = 3.27$ months), but the two gender groups did not statistically differ on motor skills ($t(30) = 1.538$, $p = .134$).

Gross motor scores did not correlate with looking time to normal actions ($r = -.051$, $p = .782$). However, the score of gross motor functions showed a marginal positive correlation with looking time to moonwalk actions ($r = 0.309$, $p = .085$), suggesting that infants with better gross motor skills were likely to show longer looking time to moonwalk actions with reversed body translation. Next, we further refined the motor assessment by focusing on the walking-related questions, because the experimental stimuli were largely related to walking actions and walking is the milestone of motor development in the age range of the infants in our study. We narrowed the analysis to the three questions highly relevant to assessing the development of walking skills: Question 4 ("when placed into a standing position, your child will stand alone for a few seconds without helping"), Question 7 ("when walking down a hallway or small room, your child will walk in a straight line with arms lowered and swinging freely"), and Question 8 ("during free play or pretend play, you notice your child is able to walk backwards for several (5 or more) steps"). We calculated the walking-related scores by summing the standardized scores for these three questions. We found that these walking-related scores were highly correlated with looking time to moonwalk actions ($r = 0.422$, $p = .016$), suggesting a strong association between the development of walking ability with the novelty preference to moonwalk actions.

As expected, age was positively correlated with gross motor skills ($r = .582$, $p < .001$) and walking-related motor skills ($r = .675$, $p < .001$), indicating that gross motor skill increased with age. Age produced a marginally significant correlation with looking time to moonwalk actions (r = .345, p = .053). A stepwise multiple regression was conducted to evaluate whether both age and walking-related motor scores were necessary to predict infants' looking time to moonwalk actions. At step 1 of the regression analysis, walking-related motor scores were selected as showing a significant relation to looking time to moonwalk actions (F(1,30) = 6.488, p = .016). We found that 42.2% of the variance of the looking time could be accounted for by more developed walking motor skills. Age did not enter into the regression equation at step 2 of the analysis (t = .492, p =.626). These regression results indicate that walking-related motor skill was the primary predictor variable of looking time to moonwalk actions.

Impact of spatial locations of the body structure on looking time on the first trial. To analyze whether infants paid more attention to some body regions than others, we defined two regions of interests (ROIs) for each action: the upper half of the screen, containing upper limb movements of the human motion, and the lower half of the screen, containing leg movements of the actions. The two hip points were aligned with the vertical center of the video at the beginning of each trial. However, hip points might travel across the upper half and the lower half of the video as the action was performed. A mixed ANOVA was conducted, with action type (normal actions with matched translation vs. moonwalk actions with reversed translation) and ROIs (upper vs. lower regions of display) as within-subject variables, performability as a between-subject variable, and looking time on the first trial as the dependent variable. The main effect of ROIs was significant ($F_{1,\,32}$ = 9.514, p = .004, η_p^2 = .849), with longer looking time at the upper half of the body (M = 3883, SD = 1653) than the lower half (M = 2470, SD = 1404). The longer looking time at the upper body may reflect an analysis of body translations, since the torso and the head are relatively rigid parts, serving as good indicators of body positions in the environment. The main effect of action type (i.e., the motion consistency constraint) remained significant ($F_{1,\,32}$ = 4.310, *p* = .046, η_p^2 = .521), as infants looked longer to moonwalk actions (i.e., reversed body translation). The main effect of performability was not significant ($F_{1,\,32}$ = 0.042, *p* = .839), nor were any interaction effects.

We further examined if the preference to the upper half of the body was related to age. Since there was no interaction between action type and the spatial location, we merged the looking times of actions with matched and reversed body translations to obtain the average looking time for the upper half and the lower half of the video. We calculated the scaled difference between looking time for the upper half versus the lower half (i.e. $\frac{(\text{Upper}-\text{Lower})}{(\text{Upper}+\text{Lower})}$) to assess visual preference for spatial locations. Age was marginally positively correlated to relative looking times for the upper and lower half of the body (*r* = 0.321, *p* = .064), suggesting that with age infants showed a stronger preference for the upper half. Walking-related motor scores did not show a significant correlation with relative looking time for the upper versus lower half (*r* = 0.238, *p* = .189).

4.4 Discussion

The key finding of the study is that by 18 months, infants can discriminate actions based on consistency between motion cues. Infants were likely to show a novelty preference in response to stimuli similar to Michael Jackson’s moonwalk dance. Furthermore, the ability to detect the violation of the motion consistency constraint was related to the development of gross motor functions, especially walking-related motor skills in the age range of 9 to 18 months. This relation implies that infants with better-developed gross motor functions are likely to show a stronger tendency to visually explore novel actions when viewing novel stimuli (in particular, the first exposure to skeleton displays of human body movements). Our results suggest that infants' visual sensitivity to subtle characteristics of human actions is intertwined with the development of gross motor skills. Within the age range of 9 to 18 months, gross motor functions develop quickly in infants. This motor experience may enable rapid developments in multiple cognitive domains (Piaget, 1953), including 3D object recognition (Soska, Adolph, & Johnson, 2010), holistic face processing (Cashon et al., 2013), and language development (Walle and Campos, 2014; He, Walle, & Campos, 2015). The present study provides converging evidence in support of the connection between visual ability in action perception and motor skills in whole body movements. Such connections may serve as a building block for the development of mirror neurons (Rizzolatti, Fogassi, & Gallese, 2001) to link observed actions with internally generated actions (e.g., Rizzolatti & Arbib, 1998; Rizzolatti & Fadiga, 1998).

Another intriguing finding is that infants showed a preference for the upper half of the moving body relative to the lower half. This result seems inconsistent with previous findings with adults showing that movements of feet carry the most important information for the perception of walking direction (Troje & Westhoff, 2006), and with a previous report that infants display far greater interest in lower and middle regions versus the top region in point-light displays depicting human biological motion (Tsang, Ogren, Peng, Nguyen, Johnson, & Johnson, 2018). However, previous studies focused on action recognition, such as recognizing walking directions or gender, based on displays showing walkers moving in place without body translation. When the body moves in the environment, the upper part of a moving body provides important information about the spatial location of the agent in the environment. The attention to the upper body observed in the present study implies that infants have a basic understanding of where to look to reliably extract the location of a moving agent. This preference exhibited in the

9-18-month olds may also result from greater experience in moving the upper body. With most infants developing the ability to crawl around 7 to 10 months and walk around 9 to 12 months, in the early infancy, interaction is dominated by upper-body movements such as reaching for parents and grasping objects. This experience with upper-body movements may drive selective attention to the upper half of displays of bodies in motion.

The motion consistency constraint results from a physical causal relation: limb movements *cause* body displacements in the environment. Sensitivity to this constraint in infants is consistent with previous infant studies using physical motion of objects (Leslie, 1982; Leslie & Keeble, 1987; Oakes & Cohen, 1990; Woodward & Sommerville, 2000). Previous research has shown that near the end of the first year, infants are sensitive to spatiotemporal features of movements of physical objects—cues related to the emergence of causal perception. Typical motion stimuli have been launching events (Michotte, 1963), in which one moving object makes contact with a static object and "launches" it. Oakes and Cohen (1990) found that starting at roughly 10 months, infants start to discriminate launching events on the basis of causality. Other evidence indicates that the development of causal perception may originate as early as 6 months (Leslie & Keeble, 1987), perhaps driven by spatiotemporal features of perceived dynamic scenes (Leslie, 1982).

Although the current study demonstrates that infants show discrimination for actions in accord with the motion consistency constraint, the results cannot differentiate between two hypotheses regarding the origin of the effect. Perhaps infants possess a general understanding of the physical laws and causality involved in human actions, or perhaps the effect is driven by specific spatiotemporal features that signal congruency in human body movements. In future studies, causal perception of human actions in infants can be examined to gain a deeper understanding of the developmental trajectory connecting causal perception with action perception.

CHAPTER 5

Revealing the Neural Mechanism of Causal Action Perception Using MEG

5.1 Abstract

Humans can effortlessly extract various information from impoverished motion stimuli such as point-light displays depicting human actions, including the basic motion information but more importantly, an understanding of goals and intentions of actions. This requires an understanding of the causal relation between limb movements and the corresponding body motion, that limb movements propel body motion to achieve certain goals. However, the neural mechanism underlying the causal reasoning in action perception remains unknown. Here, we used magnetoencephalography (MEG) decoding to investigate the spatiotemporal dynamics of neural activities associated with causal reasoning in action perception. In task 1, we decoded MEG channels that are sensitive to different action categories and we further used the selected channels in task 1 to decode causality in task 2. We found that action categories in task 1 can be decoded as early as after 200 ms regardless of specific action categories. The decoding of causality in task 2 happened later in time and heavily relied on the temporal dynamics of different actions. Source localization showed that while attention and visual related regions, including the primary visual cortex, lateral occipitotemporal cortex (LOTC), and superior temporal sulcus (STS) were strongly associated with the action classification task, a more distributed network was activated for the causal judgment task, involving orbitofrontal cortex (OFC), inferior parietal lobule (IPL), premotor cortex, dorsolateral prefrontal cortex (DLPFC), and inferior temporal gyrus (ITG).

5.2 Introduction

Recognizing human body movements is considered as one of the most sophisticated abilities supported by the human visual system. In addition to extracting basic motion information from action stimuli (e.g. speeds, moving directions), humans have been shown to be able to efficiently extract many high-level information embedded in actions, such as inferring goals of actions and social interactions even for infants who are just a few months old (Woodward, 1998; Woodward & Guarjardo, 2002). Critically, making inferences of goals of actions requires the implicit knowledge of causal constraints that human actions support the generation and the understanding of actions. For example, observing certain limb movements triggers the expectation of changes in body position. Evidence has shown that humans are sensitive to the congruency relation between limb movements and body translations. Violating the motion congruency constraint causes a significant decrease in the perception of animacy (Thurman and Lu, 2013) and social interactions (Thurman & Lu, 2014a). In addition, the congruency relation between motion cues has a causal direction with a specific temporal order, as evident in Chapter 2 (Peng, Thurman, & Lu, 2017), suggesting an important role of causal perception in the relation between limb movements and body translation. Despite the importance of this causal constraint in action perception and understanding, the neural mechanism behind the process remains unknown.

Previous studies investigating the neural mechanism of causality mostly focused on physical events such as the “launching” effect. For example, Fugelsang et al. (2005) used functional magnetic resonance imaging (fMRI) to examine the neural correlates of perceptual causality while participants were viewing alternating blocks of causal launching events and non-causal events with either a spatial or a temporal gap between the movements of the two objects. They found that causal events elicited significantly greater higher activation in the right middle frontal gyrus and in the right IPL than non-causal events, suggesting the important role of right hemisphere loci for perceiving causality for launching displays. Similarly, Wood and his colleague (2014) investigated the neural systems involved in the processing causality while different perceptual cues were involved in the launching effect. In the experiment, spatial linearity was manipulated by changing either the relative angle between two balls and temporal contiguity was varied by introducing a different temporal gap between the contact of the first ball and the onset of the second ball's movement. Results showed a dissociation between brain areas involved in processing different perceptual cues for causal perception. When using spatial information, the neural activation was increased in frontal and parietal regions including inferior frontal gyrus (IFG), right superior parietal lobule (SPL) and IPL. In contrast, when participants judged causality based on temporal information, cerebellar vermis and right hippocampus were activated. Hence, a flexible brain network is recruited to process visual information in the way of supporting causal perception.

Blos and his colleague (2012) further studied the neural processes underpinning causal perception in different contexts: physical causality and social causality. In the physical context, a classic launching event happened between a blue ball and a red ball. In the social context, the red ball was positioned above the center and the blue ball first moved horizontally and then advanced to one of the seven directions once it reached the middle. This animation was usually perceived as a social event that the behavior of the blue ball was changed due to the presence of the red ball and curved the trajectory. Results showed that the tasks in both contexts recruited similar brain areas, but elicited very different or even opposite activation patterns. For example, the activation of left insula increased for causal trials in physical contexts but not for causal social contexts. Interestingly, the neural activation for causal physical context correlated more to the activation for non-causal trials in social contexts. Furthermore, if the activation of the two contexts is mixed together, the data showed no main effect (i.e. causal physical and social contexts versus non-causal physical and social contexts) of causality, suggesting that no significant activation for causal judged stimuli contrasted to non-causal judged stimuli. This indicates that the perception of causality may not be a universal cognitive process that is implemented in certain brain regions, but, instead, depends on the context of events (i.e. physical or social).

The other line of research studies relevant to causal actions are about neural mechanisms underpinning action perception. Previous neuroimaging studies on human action recognition have provided important evidence on both the spatial locations of brain regions and the temporal dynamic of neural activities associated to processing human actions. Regarding spatial locations of brain regions, numerous studies have found body-selective neural mechanisms in the visual cortex either using single-cell recordings in primates, or through methods not limited to event-related potential (ERP), fMRI, transcranial magnetic stimulation (TMS) in human studies (for a review, see Peelen, & Downing, 2007). The results showed some consistent brain regions in visual cortex that selectively respond to body images, such as bilateral extrastriate body area (EBA) (Downing, Jiang, Shuman, & Kanwisher, 2001; Spiridon, Fischl, & Kanwisher, 2005), fusiform body area (FBA) (Peelen, & Downing, 2005; Peelen, Wiggett, & Downing, 2006; Schwarzlose, Baker, & Kanwisher, 2005), the inferior temporal cortex (IT) (Desimone, Albright, Gross, & Bruce, 1984; Gross, Bender, & Rocha-Miranda, 1969), and superior temporal sulcus (STS) (Tsao, Freiwald, Knutsen, Mandeville, & Tootell, 2003; Pinsk, DeSimone, Moore, Gross,

& Kastner, 2005). Other studies revealed brain regions that are activated when viewing human actions as dynamic displays, including posterior STS (Allison, Puce, & McCarthy, 2000; Grossman, & Blake, 2002), parietal cortex (Bonda, Petrides, Ostry, & Evans, 1996) and the ventrolateral premotor cortex (vPMC) (Buccino et al. 2004), as well as the EBA (Urgesi, Candidi, Ionta, & Aglioti, 2007). Specifically, posterior STS (pSTS) was shown to respond to a change of intentions beyond physical properties such as motion energy and saliency (Gao, Scholl, & MyCarthy, 2012).

As for the temporal dynamics of neural activates of action perception, many previous studies used EEG and MEG to reveal the time course of the visual processing of body movements with biological motions. For example, a recent study by Isik, Tacchetti, & Poggio (2018) used MEG to investigate how early it is to be able to decode actions that are invariant to viewpoints and forms of displays. They found that as early as 250ms, action categories can be decoded from MEG signals invariant to viewpoints of the motion stimuli. The results are consistent with the findings by Tucciarelli et al., (2015) showing that MEG decoding could classify action categories after 200 ms and LOTC has the earliest access to abstract action representation. Other studies used EEG to investigate the neural activities when observing biological motions and found that a negative peak at around 200 ms (N200) is associated to biological motion and the peaks were significantly larger in the biological motion condition compared to the scrambled motion condition (Hirai, Fukushima, & Hiraki, 2003) and form and motion information are therefore integrated by approximately 200 ms (White, Fawcett & Newman, 2014). Similar findings were also reported for upright vs. inverted walkers and the source was traced back to right fusiform gyrus and the right superior temporal gyrus for the second component (Jokisch, Daum, Suchan, & Troje, 2005; Virji-Babul, Cheung, Weeks, Kerns & Shiffrar, 2007).

However, few studies so far have looked beyond action classification and further into the visual processing of motion constraints that guide our understanding and predictions of actions. Here, we use MEG decoding and source localization to reveal the neural mechanism underlying causal perceptions in action, specifically the mechanisms supporting the causal constraints between limb movements and body translations in human action perception. We generated "moonwalk" actions that violate the causal expectation between limb movements and body translations. Using two tasks, we aim to compare the neural mechanism of action classification and causal perception in actions, including the time courses of causal judgment and the spatial distribution of brain activations. We hypothesize that the processing of causal perception in actions happens later in time than the time course for the action classification process, which is likely due to the need of processing time for integration of multiple motion cues. We predict that the causal process in action perception may involve more brain regions out of visual cortex, as the temporal, parietal, and frontal cortex are associated with the judgment of spatiotemporal contingency in causal actions.

5.2 Methods

5.2.1 Participants

The research was approved by the Committee for Protecting Human & Animal Subjects, School of Psychological and Cognitive Sciences at Peking University, Beijing. MEG and behavioral data were collected from 6 participants (2 females, mean age (SD) = 20.33 (1.36)) with normal or corrected to normal vision. Participants who completed the study received 160 yuan for monetary compensation. All of the participants finished an MRI scan only after the MEG session or one week before. Potential subjects were screened via a questionnaire to make sure they were eligible for MEG recording and subsequent MRI structural scans and had no history of mental illness or use of psychoactive medication.

5.2.2 Stimuli

Action stimuli were generated from the Carnegie Mellon University Motion Capture Database (http://mocap.cs.cmu.edu). We selected three action categories with clear body displacements over time: jumping, running and walking. Each action category contains 5 exemplar actions. Point-light videos were generated using the Biological Motion Toolbox (van Boxtel & Lu, 2013) with 13 main joints of the body, including the head, shoulders, elbows, wrists, hips, knees and ankles. The lower half of image frames included a checkerboard floor (see Figure 5.1). The height of skeletal actors was normalized for all actions as 8˚ visual angle. The average horizontal locations of the hip points in each video were kept at the horizontal center. Body displacements were computed as the change in the average position of the two hip joints in time, and limb movements were defined as the residual motion after subtracting body displacements on a frame-by-frame basis (Peng et al., 2017). For each of the four actions, a non-causal action was generated by reversing the horizontal direction of body displacements (see an illustration in Figure 1). The stimuli used in the experiments can be viewed at https://yujiapeng.com/megdemo.

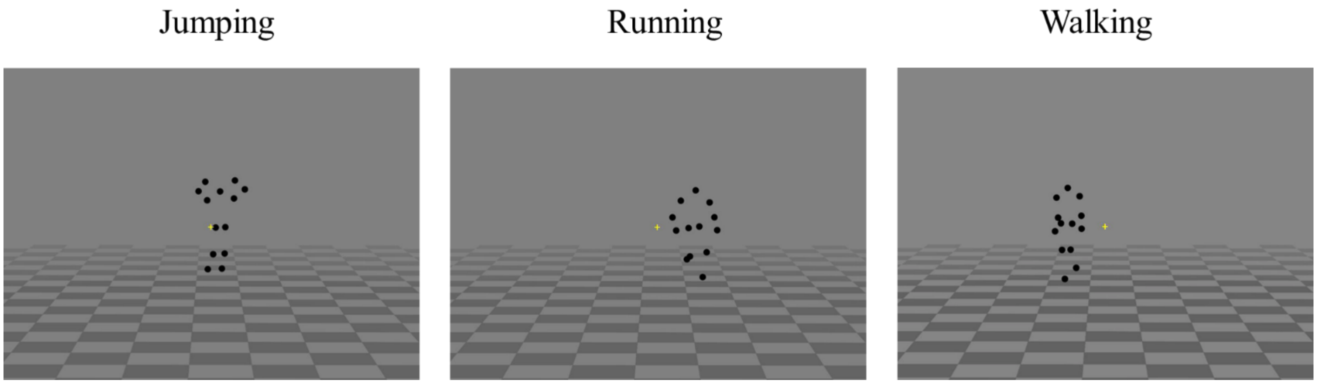

Figure 5.1. Sample frames of three types of actions. From left to right, the figures show static frames of jumping, running, and walking actions respectively.

5.2.3 Procedure

Task design. The experiment contains two phases with two distinct tasks. In phase 1, participants were told that there would be three action categories and their task was to judge the action category by pressing 1, 2, or 3 on the response box for jumping, running, and walking respectively. Participants were asked to fixate on the central cross and remain still as much as possible throughout the whole experiment. Phase 1 contains 5 blocks, each with 60 trials (i.e. 20 jumping trials, 20 running trials, and 20 walking trials), yielding a total of 300 trials. Videos were 3 s long in each trial, followed by a jittered blank post-stimulus period of 1, 1.5 or 2 s. Participants were asked to respond after the 3 s video presentation.

In phase 2, participants were instructed that there would be natural and unnatural actions displayed in this phase. Natural actions correspond to typical actions that we observe in our daily life. Unnatural actions are those that appear to violate physical laws. For example, the point-light actor may seem to be moving forward but the body is dragged by some external forces to an inconsistent direction. The task was to judge if the action in each trial looks natural or unnatural by pressing 1 or 2 correspondingly. Part 2 contains 8 blocks, each with 60 trials (i.e. natural/unnatural × jump/run/walk × 5 exemplars × 2), yielding a total of 480 trials. The entire experiment takes about an hour. No practice trials were conducted and no feedback was provided throughout the entire experiment.

Separate from the MEG experiment, a behavioral pilot experiment was conducted to collect the reaction time information with a different group of subjects. In the MEG experiment, participants were asked to respond after viewing the entire 3s to avoid contaminating MEG signal with motor signals, the reaction time cannot reflect the real speed of the judgments. To acquire more accurate response-time measures of how fast participants can respond to the two tasks, we further conducted a behavioral experiment with 20 UCLA students (mean age = 20.7, 18 females). Same stimuli were used as in the MEG experiment and the same procedure was conducted except that participants were asked to respond as fast as possible after the onset of video stimuli. Accuracy and reaction time were recorded for each trial.

MEG acquisition. MEG recordings were obtained with a 306-channel Neuromag Vectorview whole-head system (Elekta Neuromag, Stockholm, Sweden) with 204 planar gradiometers and 102 magnetometers enclosed in a magnetically shielded room (Vacuumschmelze, VAC) with a shielding factor of more than 1,000,000 at 1 Hz, with additional active shielding a shielding factor of more than 5,000,000. Data were sampled at 1,000 Hz. Before the beginning of the recording session, four head position indicator (HPI) electrodes were affixed to monitor head position in the dewar, with two attached asymmetrically to each participant’s forehead and two attached to the mastoid processes behind ears. Digitizer data were collected for each participant’s head on a Polhemus FastTrack 3D system within a head coordinate frame defined by anatomical landmarks (left preauricular area (LPA), right preauricular area (RPA) and the nasion). HPI positions were marked within this frame, and 100–200 points on the scalp and the face were saved for use in co-registering with a multi-echo structural MRI of the subject. Eye movements and blinks were monitored via two electrooculography (EOG) electrodes: one vertical electrode on the left eye, placed just above the eyebrow, and the other on the upper cheekbone just below the right eye. One electrode was placed on hand as the ground. Recordings were stored for offline analysis.

5.2.4 Data analysis

Data pre-processing and averaging. The MEG signals were first filtered using temporal Signal Space Separation (SSS) with Elekta Neuromag software MaxFilter, which aims to separate magnetic signals coming from within the brain from those coming from outside the brain in a mathematical way. Pre-processing and averaging of all recordings were performed with the minimum-norm estimate (MNE) analysis package (Gramfort et al., 2014), MNE-Python (Gramfort et al., 2013), Neural Decoding Toolbox (Meyers, 2013) and custom scripts in Python and Matlab. For all the following analysis, a low-pass filter of 60 Hz was applied, and recordings were epoched from 200 ms before stimulus onset to 3000 ms post-stimulus.

Decoding was conducted for task 1 and task 2 separately. In the decoding procedure, a pattern classifier was trained to associate the patterns of MEG data with the identity of action category (i.e., jumping, walking and running) in task 1, and with the identity signaling the absence/presence of motion congruency in the actions in task 2. The decoder with the MEG signal was evaluated by testing the accuracy of the classifier on a separate set of test data. This procedure was conducted separately for each subject and multiple re-splits of the data into training and test data were utilized. The 306 channels (including both the magnetometers and gradiometers) were used as classifier features, each containing a time series data measured over time. We averaged the data in each channel into 100-ms overlapping bins with a 10-ms step size and performed decoding independently at each time point. Decoding analysis was performed using cross-validation, where the data set was randomly divided into 5 cross-validation splits. The classifier was then trained on data from four splits (80% of the data) and tested on the fifth, held-out split (20% of the data) to assess the classifier's decoding accuracy. The whole cross-validation procedure was repeated 10 times for each subject.

Decoding: feature preprocessing. To improve the signal-to-noise ratio, we averaged trials of the same action category together. In each cross-validation split, every 4 different trials were randomly chosen from each action category for each subject and were averaged together. For example, in task 1, each action category got 100 trials. Among the 100 trials, every 4 trials were averaged, yielding 25 data points per action category. In task 2, each action category got 80 trials under both the causal and the non-causal condition. Again, every 4 trials were averaged together in a random selection manner, yielding 20 trials per action per causal condition. We next Z-score normalized that data by calculating the mean and variance for each sensor using only the training data. We then performed sensor selection using only the training data in task 1. The selected sensors are carried on to task 2 for decoding. Two ways of sensor selection were conducted. In method 1, we applied a three-way ANOVA to each sensor's training data to test whether the sensor was sensitive to the different action categories. We use sensors that were selective for action category identity, i.e., showed a significantly greater variation across class than within class, with $p < 0.05$ significance based on an F-test (if no sensors were deemed significant, the one with the lowest P value is selected). The selected channels were then fixed for each time point and were used for testing. In method 2, we conducted PCA on each timepoint with all 306 channels. The 20 top principle components (PCs) were selected and the transformation from raw channel to PCs was carried on to the decoding of task 2. To avoid circularity in our feature preprocessing, the test data was never used for the Z-scoring or feature selection.

Decoding: classification. The preprocessed MEG data were then fed into the classifier. Decoding analyses were performed using a maximum correlation coefficient classifier, which computed the correlation between each test vector and a mean training vector that is created from taking the mean of the training data from a given class. Prior work has also shown empirically that results with a correlation coefficient classifier are very similar to standard linear classifiers like support vector machines or regularized least squares (Isik et al. 2014). Each test point was assigned the label of the class of the training data with which it was maximally correlated. When we refer to classifier "training" this could alternatively be thought of as learning to discriminate patterns of electrode activity between the different classes in the training data, rather than a more involved training procedure with a more complex classifier. We repeated the above decoding procedure at each time bin to assess the decoding accuracy vs. time. We reran the above procedure 10 times for each subject. We measured decoding accuracy as the average percent correct of the test set data across all decoding runs and reported decoding results for the average of 6 subjects in each experiment.

Significance testing. All statistics computed using non-parametric cluster-level one sample t-tests. The procedure determines significance through 5000 permutations with a critical α-value of

0.05, following Maris and Oostenveld (2007). Cluster mass was determined by summing statistical values of adjacent suprathreshold timepoints within a cluster. The null distribution was built by permuting data and storing the largest cluster mass observed in the permuted data. For each experiment and decoding condition, t-statistics were computed at each time point. The supra-threshold time courses of t-statistics were grouped into clusters and were compared to a null distribution formed in permutation. Permuted data were generated by randomly flipping the sign of data. The largest cluster from the permuted time course was entered into the null distribution and the procedure was repeated for 5000 times. A cluster was deemed significant if the proportion of cluster masses from the null distribution that exceeded the observed cluster mass was smaller than the critical alpha-level (i.e. 0.05, two-tailed). We define the decoding "onset time" as the onset of the earliest significant cluster. This provided a measure of when significant decodable information was first present in the MEG signals and is a standard metric to compare latencies between different conditions (Isik et al. 2018; Cichy et al. 2016).

Source localization. Individual participants' anatomical brain images for source localization of MEG activity were obtained with multi-band structural MRIs on a 3 T Siemens Prisma system. Brain images were reconstructed and triangulated brain surfaces generated via the watershed algorithm with the Freesurfer analysis package (Fischl et al., 2004). A decimated dipole grid was fitted to the inflated white matter surface in the shape of an icosahedron recursively divided five times to generate a 5124-point grid. One surface source space of the whole cortical surface was created for each participant based on the dipole grid described above. A forward solution was then calculated using the geometry-dependent solution calculated from the single-compartment boundary-element model. Sources closer than 5 mm to the inner skull surface were omitted from the forward solution in all cases. The MRI-head coordinate transformation for each subject was supplied to the forward model by aligning the digitizer data obtained in the original recording session (see MEG acquisition) with a high-resolution head surface tessellation constructed from the MRI data. The inverse operator was prepared with a loose orientation constraint parameter of 0.2 in order to improve localization accuracy (Lin, Belliveau, Dale & Hämäläinen, 2006). A depth-weighting coefficient of 0.8 was also set for the inverse operator to lessen the tendency of MNEs to be localized to superficial currents in place of deep sources. MEG data were source localized onto the entirety of each source space using a λ2 regularization parameter based on Signal-to-Noise Ratio (SNR, set to 3) equal to 1/(SNR^2). Evoked cortical activation was quantified spatiotemporally by taking only the radial component from a three-orientation source (x y z) at each vertex of the triangulated source space in the form of dynamic statistical parametric maps (dSPMs) based on an inverse solution regularized with an SNR of 3. DSPMs are a statistical representation of significant activity from each source per time point calculated by noise normalization on the estimated current amplitude (MNE) of a given source according to noise covariance between sensors calculated during a baseline period of 200 ms pre-stimulus (Dale et al., 2000). The noise covariance estimation model was selected automatically according to the rank for each participant (Engemann and Gramfort, 2015).

Five out of six participants were included in the source localization analysis. One subject had to be removed due to that a boundary-element model could not be computed for a signal dropout in the bottom slices of the T1 MRI potentially due to missing data of the MRI data. For both task 1 and task 2, the time courses of source estimates were computed for each individual actions. Specifically, for task 1, the averaged source estimate was computed by averaging across three action categories to reveal brain regions that are activated by observing actions in general.

For task 2, a contrast of source estimates was computed by subtracting non-causal condition from the causal condition in the source space for the three actions separately and then averaged across actions. The averaged source estimates of individual subjects were then morphed to a common reference source space.

5.3 Results

5.3.1 Behavioral results

From the behavioral pilot experiment, reaction time (RT) of correct trials and accuracy were computed for each participant for task 1 and 2 respectively. Task 1 yielded an accuracy of 0.96 (*SD* = 0.05). Task 2 yielded an accuracy of 0.92 (*SD* = 0.09). The average RT was 1.29s (*SD* = 0.23s) for task 1 and 1.58s (*SD* = 0.33s) for task 2. Reaction time was also calculated for the three actions separately for task 1 and 2. As shown in Figure 5.2, three actions yielded different reaction times in both tasks. A repeated measure ANOVA was conducted with two within-subject variables: task with 2 levels (i.e. task 1 vs. task 2), and action categories with 3 levels (i.e. jumping, running, walking). We found a main effect of tasks ($F(1,19) = 32.38$, $p < 0.001$), with task 2 getting significantly longer reaction time than task 1. The main effect of action category was also significant ($F(2,18) = 291.86$, $p < 0.001$), with jumping yielding significantly longer reaction time than the other two action categories ($p < 0.001$). The interaction effect between tasks and actions was also significant ($F(2,18) = 26.18$, $p < 0.001$).

The results show that the task of causal judgment in task 2 took longer time in general compared to the action classification judgment in task 1. Also, the reaction time of judgment depends on the type of action, due to the difference between temporal dynamics of actions.

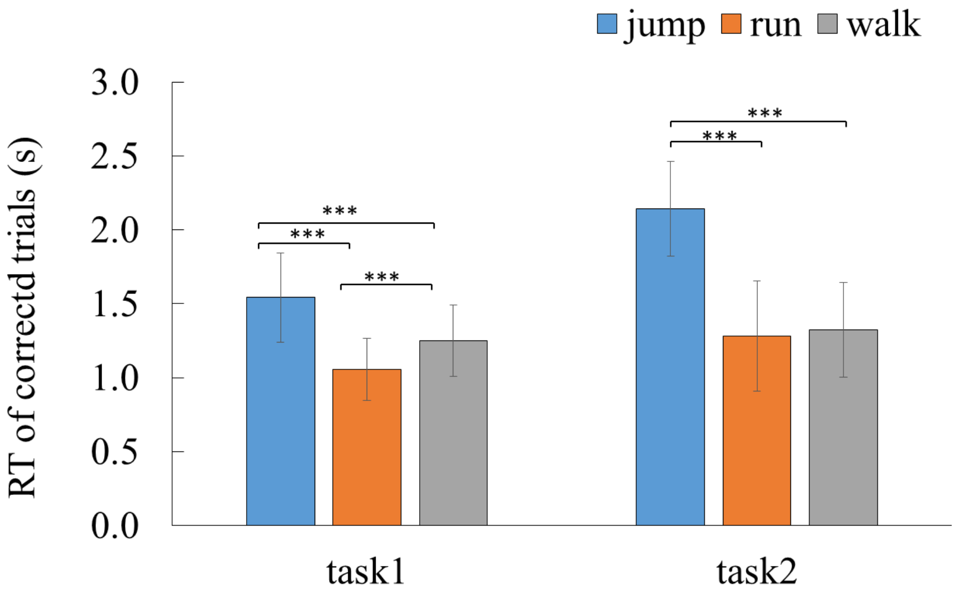

Figure 5.2. Averaged reaction times of trials with correct responses in task 1 and task 2 for the three action types respectively in the behavioral experiment.

5.3.2 MEG results

All statistics were computed using nonparametric cluster-level permutation tests based on 3000 permutations with a critical α-value of 0.05, following and Oostenveld (2007). Cluster mass was determined by summing statistical values of adjacent suprathreshold time points within a cluster rather than counting the number of adjacent suprathreshold time points (or any other cluster-weighting method). The null distribution was built by permuting data and storing the largest cluster mass observed in the permuted data.

Event-related field. The averaged event-related field (ERF) was calculated by averaging the preprocessed MEG signal for each condition for task 1 and task 2 respectively. Results of task 1 and task 2 were shown in figure 5.3a and the averaged ERFs of actions separately in task 2 were shown in figure 5.3b. No significant difference was found between action categories in task 1 in any of the four brain regions. No significant difference was found between causal and non-causal conditions in task 2 in any of the four brain regions either, even though the ERF started to show a difference between two conditions at about 300 ms in the frontal region. This difference was primarily driven by the running action.

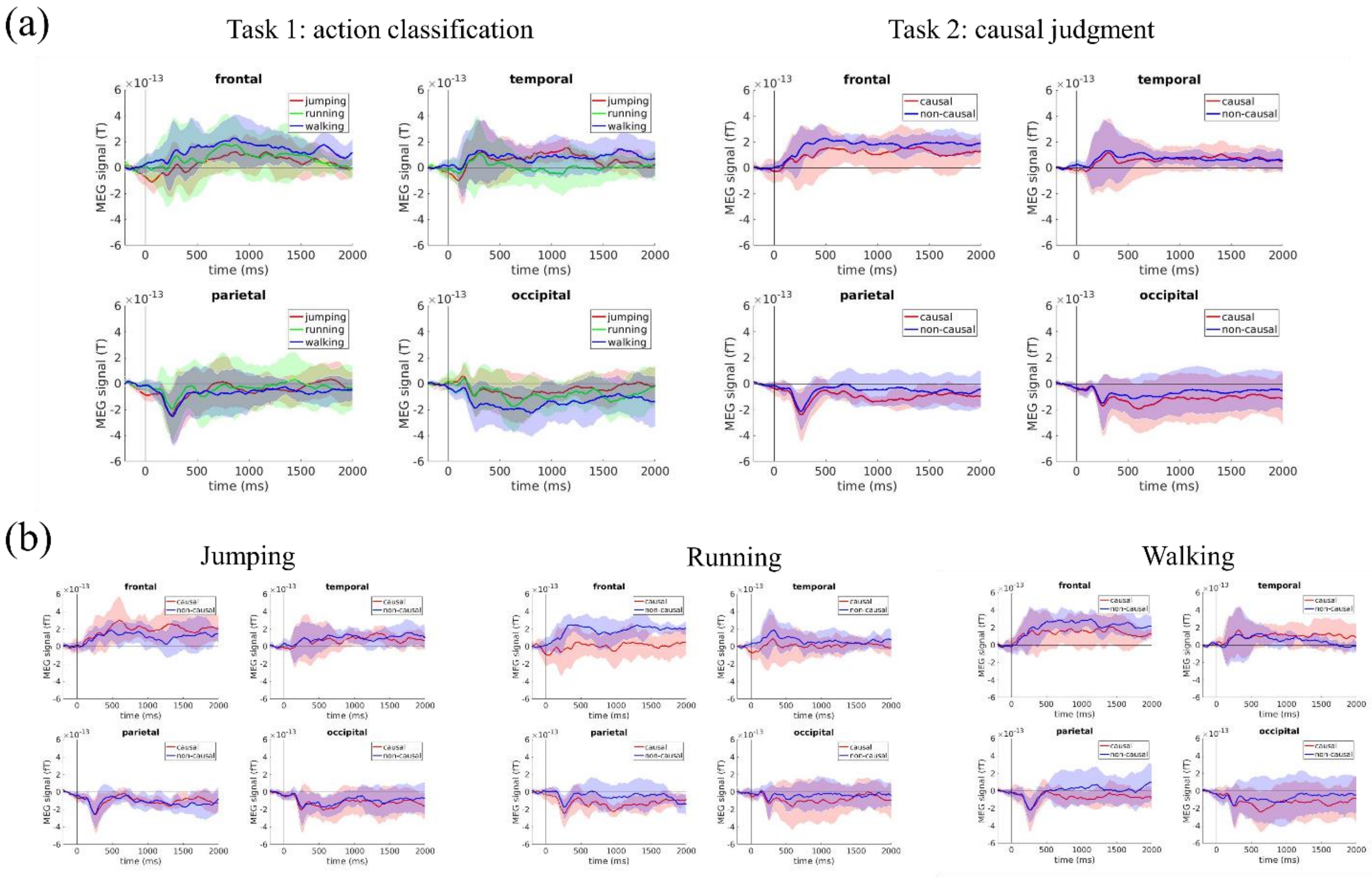

Figure 5.3. (a) Event-related field (ERF) of task 1 and task 2 for the four brain regions. (b) ERF of task 2 for the three actions separately for the four brain regions.

Feature selection. We first investigated the number of channels being selected for each task. Here, only the number of selected channels from the ANOVA feature selection method is reported. To investigate the spatial distribution of the selected channels, the 306 sensor channels are divided into eight brain regions: frontal lobes (left and right), temporal lobes (left and right), parietal lobes (left and right), and occipital lobes (left and right) (for an illustration, see Fig 1 in Hu, Yin, Zhang, & Wang, 2018), corresponding to the standard brain regions provided by Elekta Neuromag MEG. At each time point, channels that are selected from the ANOVA feature selection procedure were classified into one of the 8 brain regions. As shown in Figure 5.4, task 1 yielded a peak of number of selected channels very early, at around 200 ms. The averaged number of selected channels were significantly greater within the 200 – 600 ms window for the occipital region than the frontal region ($t(5) = 4.33$, $p = 0.008$). The comparison between occipital and temporal ($t(5) = 1.94$, $p = 0.110$), and occipital and parietal ($t(5) = 2.18$, $p = 0.082$) were not significant, but showed a trending significant difference. This suggests that occipital region played a relatively dominant role compared to parietal, temporal and occipital regions. The number of selected sensors were slightly more in the right hemisphere than the left hemisphere but the difference did not reach a statistical significance. In contrast, task 2 reached a peak of number of selected channels later in time, starting to increase rapidly after 1s. Here, the occipital region no longer has significantly greater number of selected channels compared to other cortices early in time as in task 1 ($p > 0.2$). However, the occipital region yielded significantly more selected channels between 1200 ms to 1700 ms (occipital vs. frontal, $p = 0.066$; occipital vs. parietal, $p = 0.038$; occipital vs. temporal, $p = 0.016$). This suggests that the occipital area may still greatly contribute to the causal judgment task, but the process happens later in time. Again, no difference was found between left and right hemisphere. As shown in figure 5.5, the number of channels being selected separately for three actions showed different patterns of distributions. While walking reached a peak before 500 ms, running reached a peak after 1s, and jumping reached a peak after 1.5 s. The results indicate a difference between time windows of successful decoding for three actions, probably driven by differences in the temporal dynamics of three actions. For example, the causal congruency between motion cues in walking actions may be able to be revealed early in time based on low-level visual information, while running and jumping require a longer time for cumulating visual information. The process of causal judgment may also happen early for walking driven by familiarity due to repeated exposures in daily life.

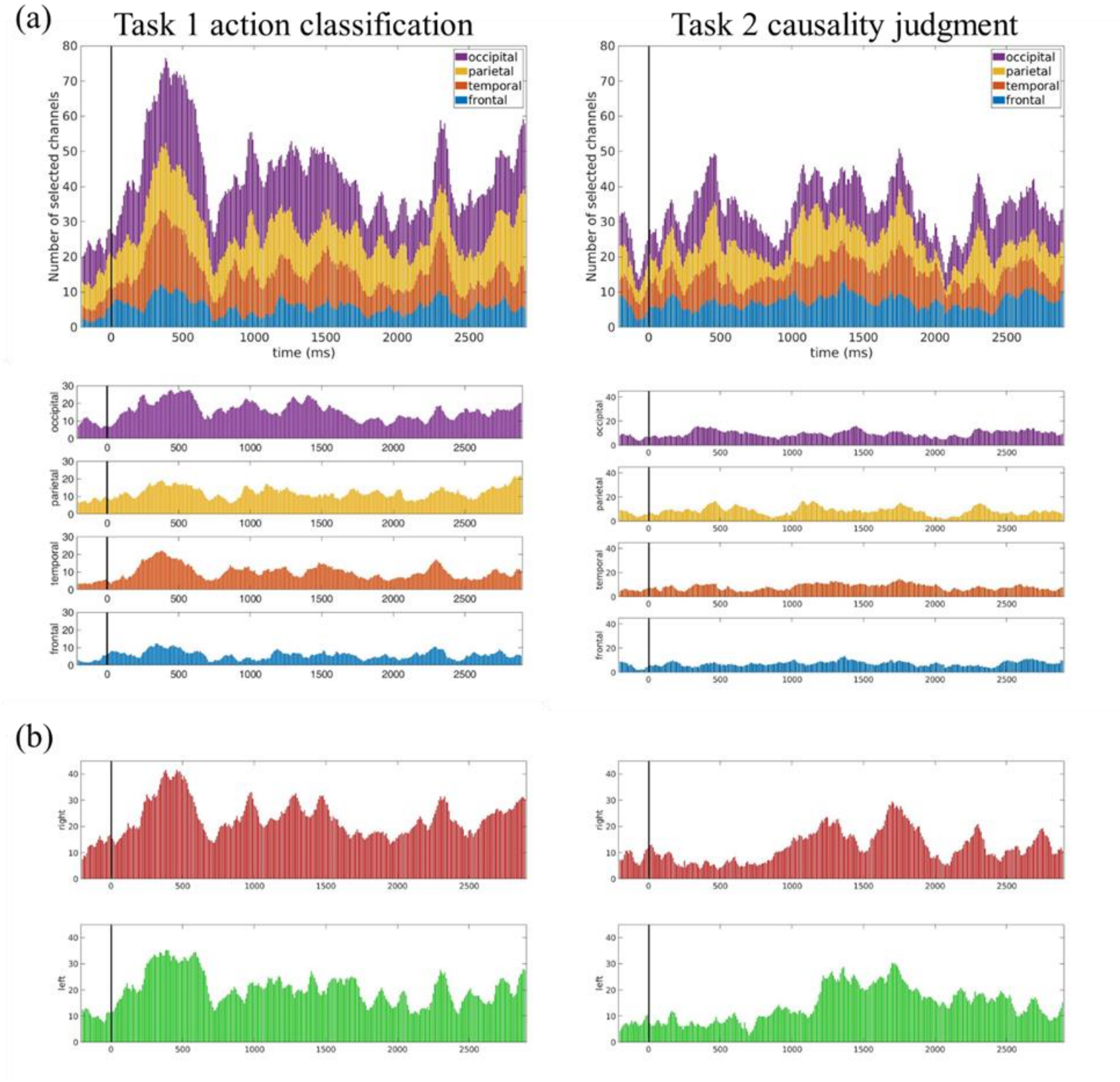

Figure 5.4. Channels being selected from ANOVA feature selection were grouped into one of the eight cortical regions. Here shows the distribution of numbers of selected channels at different time points averaging across all the subjects. (a) The distribution of numbers of selected channels from the four cortical regions, merging across left and right hemisphere. (b) The distribution of numbers of selected channels from the left and right hemisphere.

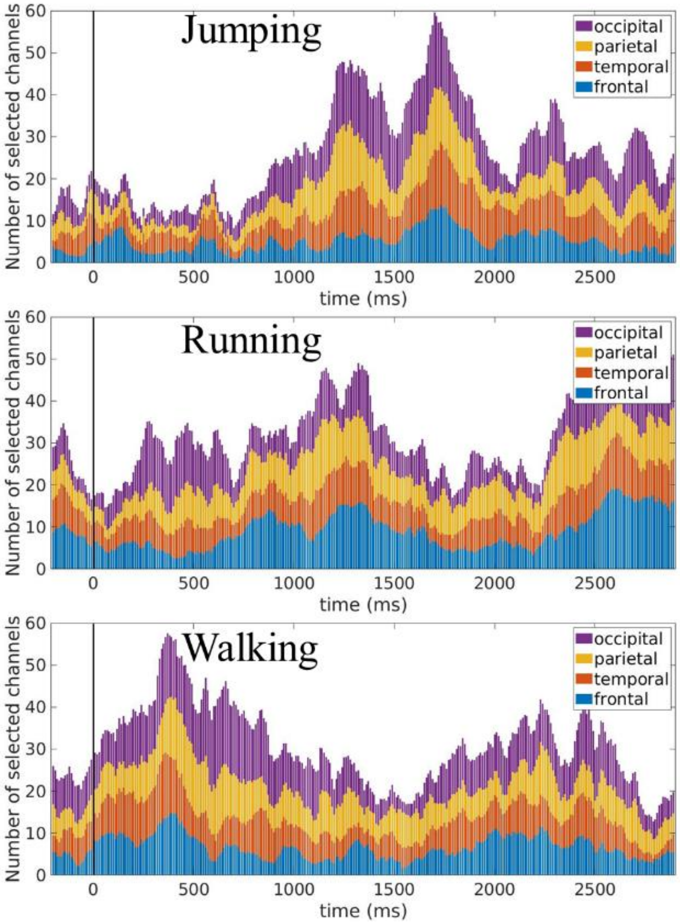

Figure 5.5. Channels being selected from ANOVA feature selection for task 2 with three actions being decoded separately. From top to bottom, it shows the distribution of jumping actions, running actions, and walking actions.

Decoding performance. Decoding results based on two methods of features selections were shown in Figure 5.6. The ANOVA feature selection method and the PCA feature selection method yielded similar decoding performance, but PCA showed slightly smoother and higher decoding accuracy. Thus, the following reports were mainly based on the PCA feature selection method. For task 1, we were able to decode action categories (i.e. jumping, running, or walking) as early as 50 ms, and the first significant chunk of time with above-chance accuracy lasted from 50ms to 810 ms, yielding an average accuracy of 0.47 within this time range, which is significantly above the 0.33 chance level. The performance reached a peak at 420 ms with an accuracy of 0.55. When the results of three actions were shown separately, they all showed a rapid increase of decoding accuracy right after 200 ms (i.e. decoding latencies were 350 ms for jumping, 240 ms for running, no significant chunk of time was found for walking but the peak was at 410 ms). The decoding curves showed two significant chunks of decoding accuracy in the jumping action condition, one before 1 s and one after 1s. This was probably due to a difference in the temporal dynamics of jumping actions compared to running and walking. While running and walking show signature movements continuously across the entire videos, the signature movements of jumping appear in the middle of the 3s video (i.e., around 1.5 s), after a short period of preparation for jumping movement.

Compared to the relative performance in task 1 (i.e. accuracy relative to the chance level), task 2 yielded slightly lower accuracy in general, with a peak accuracy of 0.60, indicating that decoding causality is a more difficult task than decoding action categories. Also, the first significant peak happened slightly later in time, at around 300 ms for the ANOVA feature selection method and after 1.5 s for the PCA feature selection method. In contrast to decoding performance in task 1, the decoding of the three actions separately in task 2 reached peak performance at dramatically different time points. Jumping yielded a decoding latency of 1580 ms, and a peak accuracy of 0.70 at 1700 ms, probably because the difference between causal and non-causal jumping actions was hard to be revealed before the signature movements of jumping. Running yielded a decoding latency at 1050 ms, with a peak accuracy of 0.65 at 1140 ms. Walking got a decoding latency of 230 ms and reached a maximum accuracy of 0.70 at 630 ms.

This suggests that the processing of causality in human actions happens after action recognition and strongly depends on the temporal dynamics of actions.

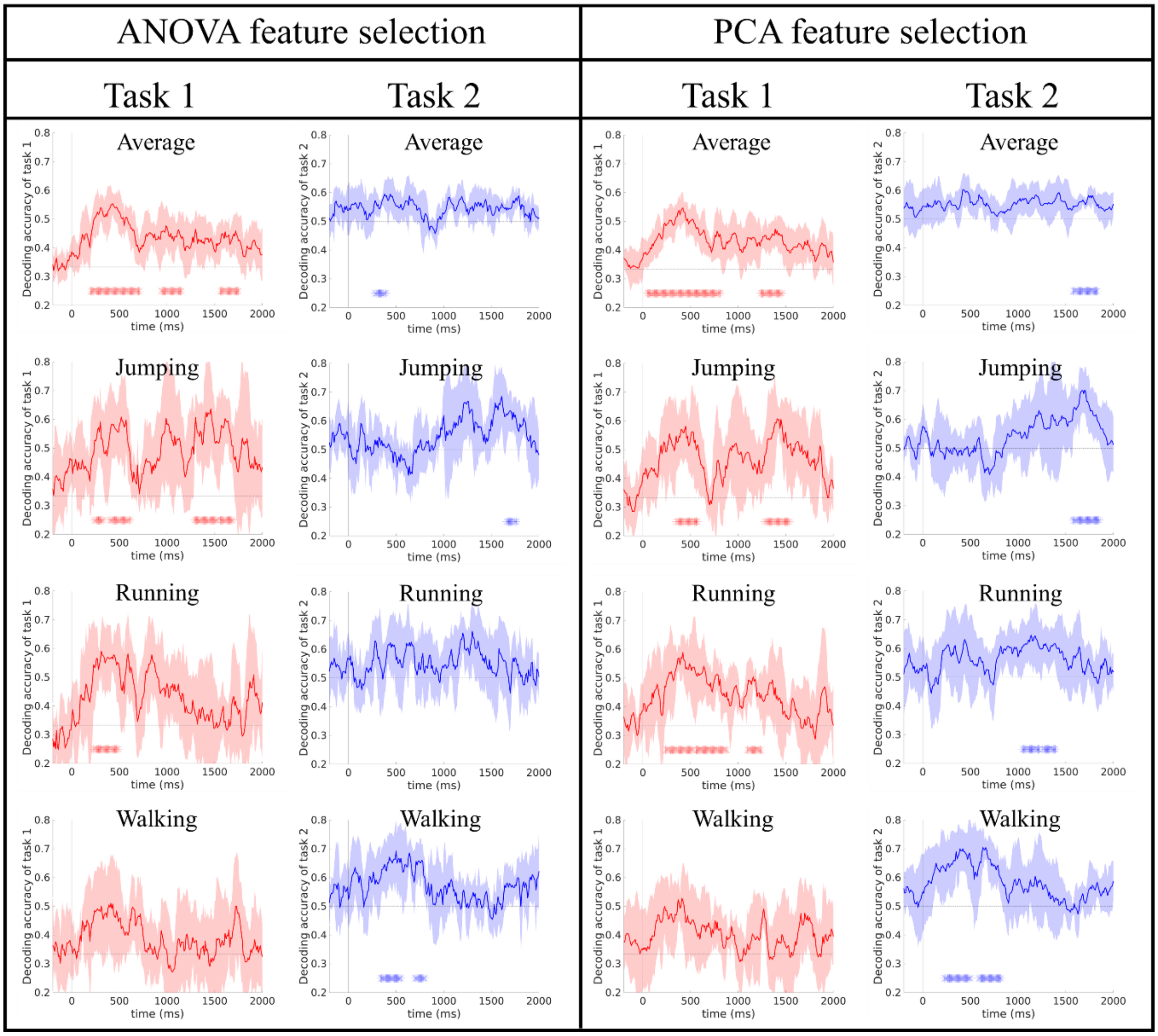

Figure 5.6. Decoding accuracy of task 1 and task 2 under two feature selection methods. The left half presents the decoding accuracy of the ANOVA feature selection method. The right half presents the decoding accuracy of the PCA feature selection method. Here, task 2 was decoded from channels that were selected entirely based on task 1. Stars indicate statistical significance calculated from a permutation one-sample t-test compared to the chance level.

Source localization. Snapshots of the spatiotemporal activity estimates of task 1 were shown in Figure 5.7 (left). Four different latencies were picked based on the decoding results in Figure 5.6. In task 1, a clear activation of the primary visual cortex was observed at 200 ms for both left and right hemisphere. The MT region and pSTS regions were activated at around 300 ms with a right hemisphere dominance. At 1000ms, activations were observed at the orbitofrontal cortex (OFC), indicating a potential decision-making process. As shown in Figure 5.7 (right), the three actions showed highly consistent spatiotemporal activity patterns.

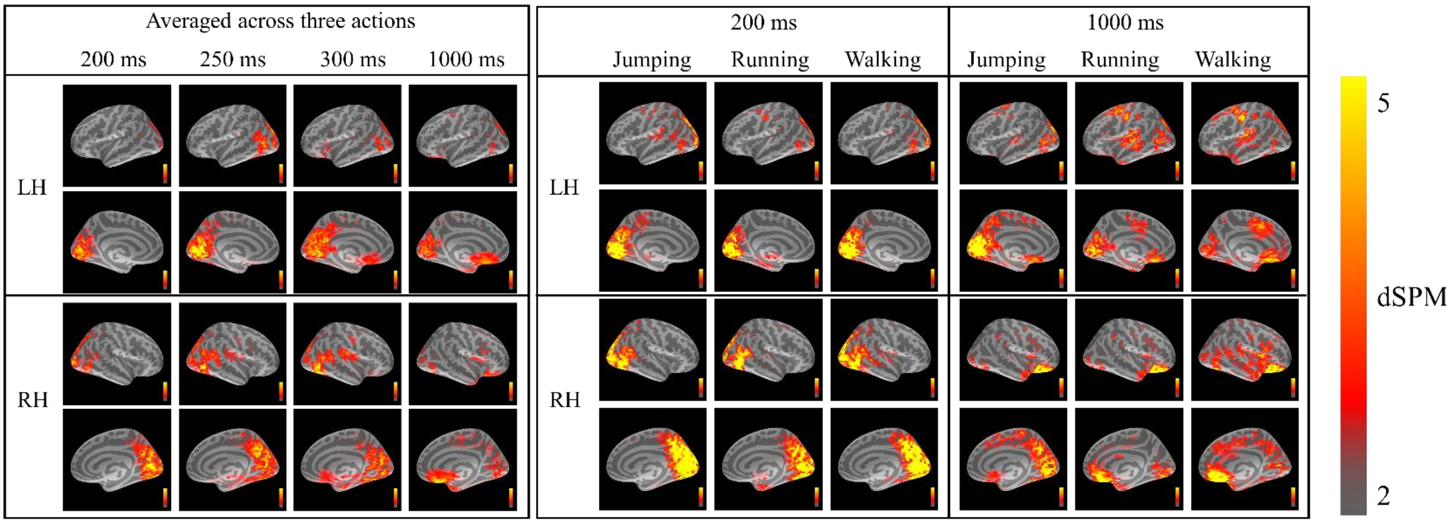

Figure 5.7. (left) Snapshots of the spatiotemporal activity estimates of task 1averaged across all action categories at four different latencies. (right) Snapshots of the spatiotemporal activity estimates of task 1of three action categories separately at two different latencies. From top to bottom the four rows correspond to left hemisphere lateral view, left hemisphere medial view, right hemisphere lateral view, and right hemisphere medial view.

As shown in Figure 5.8, a more spread pattern of activation was observed for task 2. In contrast to task 1, the primary visual cortex did not show many activations when contrasting causal from non-causal actions. The middle temporal gyrus (MTG) and ITG showed a negative contrast and the OFC showed positive contrast at 300 ms. At 1658 ms, IPL and pSTS showed negative contrasts. Due to the difference between temporal dynamics of three action categories, figure 5 (right) showed the activity estimates for three actions separately at different latencies. Jumping showed stronger activation at the middle frontal gyrus, dorsolateral prefrontal cortex (DLPFC), OFC as well as the primary visual cortex. Running showed stronger activation at IPL and superior temporal regions at 1083 ms. Walking showed strong activation early in time at around 540 ms at the premotor cortex, IPL, ITG, inferior cingulate, and superior frontal gyrus.

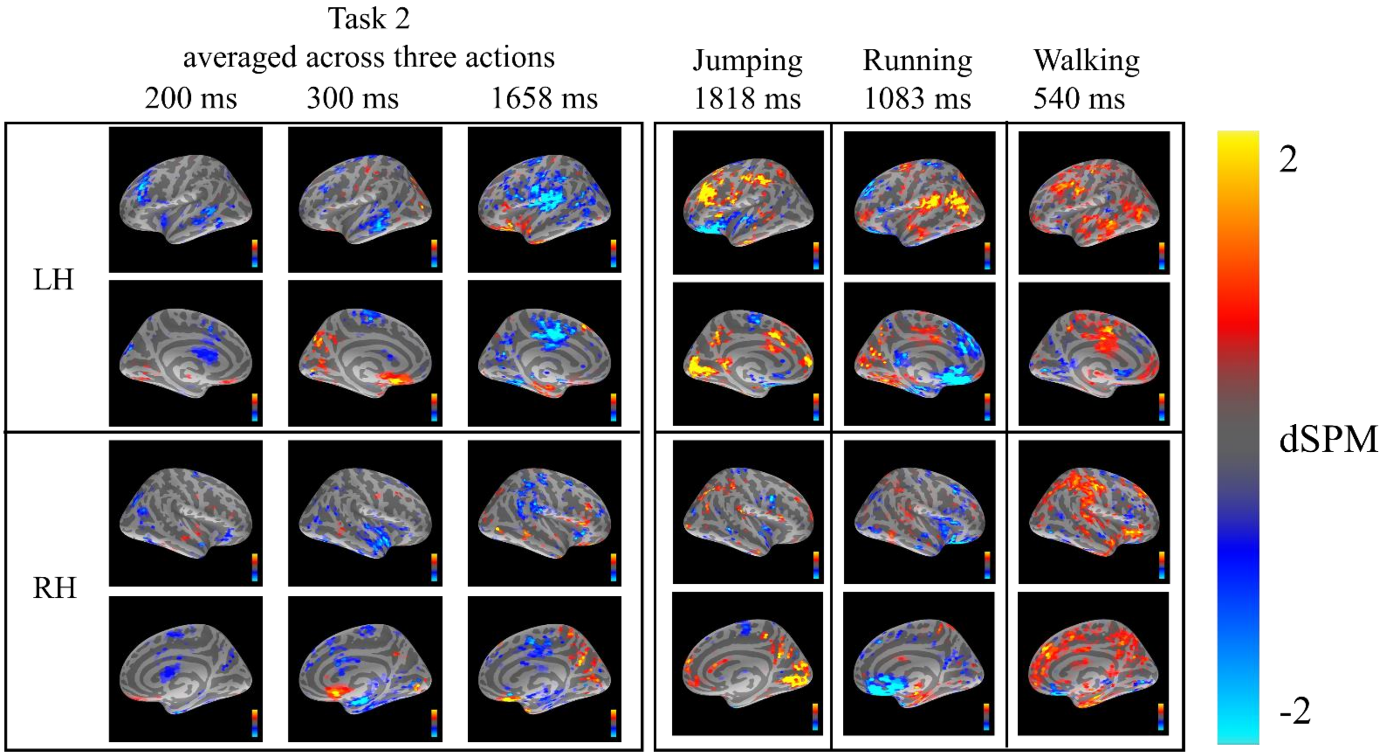

Figure 5.8. (Left) Snapshots of the spatiotemporal activity estimates of task 2 at three different latencies. (Right) Snapshots of the spatiotemporal activity estimates of three actions separately

for task 2. From top to bottom the four rows correspond to left hemisphere lateral and medial views, right hemisphere lateral and medial views. The color indicates the contrast between causal and non-causal conditions, with yellow being positive and blue being negative.

5.4 Discussions

In the current study, we used MEG to investigate the spatiotemporal dynamics of the neural representation of causal human actions and found that causality of human action can be decoded from MEG signals as early as 300 ms following the video onset, which is considerably less than the 3-s video stimuli. Substantially, the decoding was achieved by selecting channels that are sensitive to action categories in a different task, ruling out the possibility of the double-dipping. Interestingly, the causal judgment task showed dramatically different temporal dynamics of the decoding performance as well as spatial distributions of selected channels and source estimates compared to the action recognition task.

In the time domain, the two tasks reached significant decoding performance at different time points. The action classification task (i.e. task 1) got the maximum number of channels that are sensitive to action categories early in time as well as a peak in decoding performance starting at 200 ms. This is consistent with previous EEG and MEG evidence that actions can be decoded after 200 ms to 250 ms after the video onset (Hirai, Fukushima, & Hiraki, 2003; Isik et al., 2018; Tucciarelli et al., 2015). The significant decoding performance lasted form 200 ms to 700 ms, suggesting that the processing of biological motions may extend across several stages. Similar ideas were also suggested by previous EEG studies that while an early 200 ms component may be triggered by attentional effects associated with stimulus onset, a later component happens after 300 ms with specific processing of patterns of motions (Jokisch, Daum, Suchan, & Troje, 2005). In contrast, the causal judgment task reached the maximum number of channels that could differentiate causal from non-causal actions at a much later time close to 1 second after the video onset. The decoding latency of task 2 was also later in time and showed a distribution that was more evenly distributed over time and highly depended on the temporal dynamics of actions. The results suggest that the causal judgment is not temporally restricted within a time window but is a continuous process which happens along with the integration of motion cues in actions.

The two tasks also revealed highly different source estimates. The source estimates of the action classification task showed a strong activation at the primary visual cortex, precuneus, STS, as well as LOTC. The activation pattern emerged as early as 200 ms and the activation gradually spread to more anterior regions, suggesting a gradually deeper representation of motion information over time. The results are consistent with previous evidence by Tucciarelli et al., (2015) that LOTC has the earliest access to abstract action representations, between 200 ms to 300 ms. OFC was activated later in time at around 1 second, potential indicating a decision-making stage or an inhibition of motor responses (Decety, & Grèzes, 1999). In comparison, the causal judgment task showed a more spread pattern of activation both spatially and temporally. The primary visual cortex no longer shows a dominant role. Instead, MTG, ITG, IPL, DLPFC, and OFC showed activations for the contrast between causal and non-causal actions. At 300 ms, MTG and ITG showed a negative activation and the OFC showed positive activation. At 1658 ms, the IPL, inferior cingulate, and the superior temporal regions showed negative contrasts.

The results are consistent with previous evidence from fMRI on the perceptual causality in launching effects showing that right superior and IPL might contribute to causality given of their role in spatial attention (e.g., Singh-Curry and Husain, 2009; Woods et al., 2014) and spatial relational processing (Ackerman and Courtney, 2012). Left IPL activity might integrate spatial and temporal information as shown in a collision judgment paradigm (Assmus et al., 2003). IFG and DLPFC may play an important role in perceptual decision-making, category selection, and response inhibition (e.g. Heekeren, Marrett, & Ungerleider., 2008). It has also been found that bilateral IFG, bilateral IPL, and right SPL (Woods et al., 2014) were activated when participants were instructed to explicitly use spatial information to make causality judgments. In addition, the activation at premotor cortex region is consistent with previous evidence that specific causal properties such as spatiotemporal contingencies were found to be processed in monkey premotor area F5 (Caggiano et al., 2016). Sack (2009) proposed that premotor cortex may integrate sensory information and relay the information to medial-dorsal parietal cortex in a top-down manner to facilitate the spatial cognition. Meanwhile, neural inputs from premotor cortex are also sent to bilateral prefrontal cortex and occipito-temporal cortex.

Together, the current study for the first time revealed the spatiotemporal neural activities underlying the visual processing of causal constraints in biological motion. Our findings suggest a distributed neural network across frontal, parietal, temporal and occipital cortex which may be responsible for processing and integrating spatial and temporal elements over time in the dynamic causal events (Straube, & Chatterjee, 2010). However, with the current evidence, we cannot rule out the possibility that the causal judgment is based on the detection of incongruency between motion cues without referring to a high-level reasoning process. Future studies could be done with more controlled spatial or temporal elements as well as controlling for eye movements. In addition, it will be worth investigating the neural representation of causal human actions in a more natural and complex scenario such as goal-oriented action demonstrated by human-object interactions.

CHAPTER 6

Causal Action Perception in Deep Neural Networks

6.1 Abstract

Perceiving and understanding human actions are sophisticated processes achieved by the human visual system. In recent years, computer vision has developed powerful deep convolutional neural networks (CNN) that can reach human-level performance on many visual tasks, including action recognition and pose prediction (e.g. Simonyan and Zisserman, 2014; Martinez, Black, & Romero, 2017). However, it remains unknown if such neural networks acquire similar action representations as humans, and learn to achieve a deeper understanding of causal relations prevalent in human actions. Here, through a series of simulations, we systematically tested the performance of deep neural networks on the generalization of action perception and causal action understanding.

We adopted a two-stream CNN with a spatial pathway specialized in processing appearance information and a temporal pathway specialized in processing motion information. In simulation 1, we tested whether the two-stream CNN could generalize across different types of body movement displays to demonstrate the flexibility as human observers. We trained the model with actions displayed in the skeletal form and tested its performance with point-light (PL) displays. We found that only the motion stream can recognize PL actions with above chance-level performance, while the appearance pathway failed the recognition task. In simulation 2, we tested the model performance for moonwalk actions which include inconsistent limb movements and body displacements. We found that the two-stream CNN was not able to distinguish moonwalk actions from normal actions, suggesting a lack of understanding to connect certain limb movements with expected body displacements. In simulation 3, a long

short-term memory (LSTM) recurrent neural network was included in model architecture to examine if the representation of the global motion pattern is the key to reach a causal understanding of human body movements. Results showed that while the added LSTM module enabled the model to perceive global body displacements, the model failed to identify whether or not the walking action is consistent with causal relations between limb movements and body displacements. Together, we conclude that the current neural networks lack the representation for causal actions and future improvements are needed.

6.2 Introduction

Recognizing human body movements is viewed as one of the most sophisticated abilities supported by the human visual system. In daily life, we can readily recognize actions despite changes in body forms, appearance (e.g, different clothing), viewpoints, scales, and occlusions. Even for highly impoverished and rarely observed displays such as point-light displays (Johansson, 1973), in which a few disconnected dots depict joint movements, the human visual system can recognize actions with high accuracy despite visual noise (Lu, 2010), and perceive social interactions (Thurman & Lu, 2014a) and causal intention (Peng, Thurman & Lu, 2017).

In contrast to the remarkable human ability to recognize actions from body movements, developing computer algorithms to recognize action has been challenging. One representative model of action perception is by Giese and Poggio (2003), who developed a neurophysiologically plausible and quantitative model with two parallel processing streams: a ventral pathway and a dorsal pathway. The ventral pathway is specialized for the analysis of form in static frames. The dorsal pathway is specialized for optic-flow/motion information. Both pathways comprise hierarchies of neural feature detectors that extract form information or optic-flow information with increasing complexity as the hierarchy going up. Despite the fact that the model can address many neurophysiological and psychological findings, it is developed based on some assumptions that are hard to experimentally verify, and shows restricted capacity in modeling a limited set of human actions. One of the drawbacks is that the model only includes feedforward connections for both pathways, without top-down influence from other brain regions beyond visual areas. This architecture makes the model efficient in perceiving motions based on low-level motion inputs, but limits its capability under circumstances that require attention shifts and prior knowledge such as physical laws and biological structure of human bodies. Another limitation was that the model was not capable of recognizing actions from natural videos in dynamic scenes but only focusing on highly controlled biological motion stimuli, e.g, point-light display.

In recent years, with the rise of convolutional neural networks (CNNs), video recognition research has been advancing. For example, Simonyan and Zisserman (2014) extended deep convolutional neural networks (ConvNets) (LeCun, Bottou, Bengio, & Haffner, 1998; Krizhevsky, Sutskever, & Hinton, 2012), and developed two-stream CNNs which include the spatial and temporal pathways. A spatial CNN processes appearance information and a temporal CNN processes the optical flow information. At a later stage, the feature derived from the two information sources are integrated, so that both appearance and motion information is used to recognize actions. The model performed well on action classification of two challenging datasets: UCF-101 which includes 13320 videos covering 101 action types (Soomro, Zamir, & Shah, 2012), and HMDB-51 which includes 6766 videos covering 51 action types (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011). The two-stream CNNs achieved the accuracy levels in the range of 55% to 70% for the HMDB-51 dataset (compared to a chance level of 2%). This performance represents a significant improvement over previous action recognition models based on deformable templates, or part-based approaches.

Despite the success of deep neural networks in recognizing actions, it remains unknown whether neural networks trained with natural videos yield representations of actions similar to those acquired by the human visual system. In addition, whether the action representations learned in deep neural networks encompass deeper understanding about actions still remains unknown. Here, we aim to answer a few questions: (1) whether CNNs are as robust as humans in recognizing actions in novel displays, such as point-light stimuli. In simulation 1, we aim to examine the generalization of two-stream CNNs across action displays. If the network can generalize its performance to recognize human actions in the impoverished formats commonly displayed in psychophysical studies, we can also assess the contributions made by the separate pathways processing appearance and motion information, as well as contributions made by the integration of the two processes; (2) whether deep neural networks possess an understanding of the implicit causal structure embedded in human actions. In simulation 2 and 3, we used novel actions such as moonwalk action, backward action, or in-place actions, to systematically test the performance of the standard two-stream CNN, and an extended version of CNNs on causal action understanding.

6.3 Simulation 1

Skeletal and point-light displays are rarely observed in natural environments but commonly used in psychophysical research. Human adults showed a strong ability to recognize point-light displays without previous exposure. In simulation 1, we aim to address the two following questions: (1) Could a two-stream CNN reach the human-level performance of recognizing skeletal displays after training with videos showing skeletal displays of human actions from different viewpoints (2) Could the two-stream CNN model trained with skeletal displays of human actions generalize its performance to recognize actions in point-light displays. We hypothesize that if the two-stream CNN model emulated human action processing, it would be able to recognize point-light displays after training with natural and skeletal displays.

6.3.1 Model structure

The two-stream CNN relies on processing two types of information to classify a video clip into action categories. One source of information is the appearance in static image frames, and the other is motion (usually represented by optical flow fields, i.e., the spatial displacement of each pixel between adjacent frames; Horn & Schunck, 1981). This two-stream architecture is consistent with neurophysiological evidence that action processing involves both ventral and dorsal pathways in the brain, and integrates the information at action sensitive regions in the temporal lobe (e.g., Giese & Poggio, 2003).

For each of the streams in the model, we adopted the architecture of CNNs such as VGG16 (Simonyan & Zisserman, 2015) shown in Figure 6.1. The appearance stream, termed as spatial CNN, uses the input of the three channels of an RGB image; and the motion stream, termed as temporal CNN, uses a stack of optical flow vector fields spanned over a temporal window with some consecutive frames. We use 10 frames for all simulations for this temporal window. In each of the CNNs, five convolutional layers for feature extraction are applied to the input, which are followed by three fully-connected (FC) layers. If we have a total of K action categories, then the final layer (i.e. a softmax layer) will yield a vector of length K, in which the k-th element is the probability of the k-th action category being activated by the input (i.e., confidence). In the present paper, we use these two streams of CNNs to model the spatial pathway and the temporal pathway respectively.

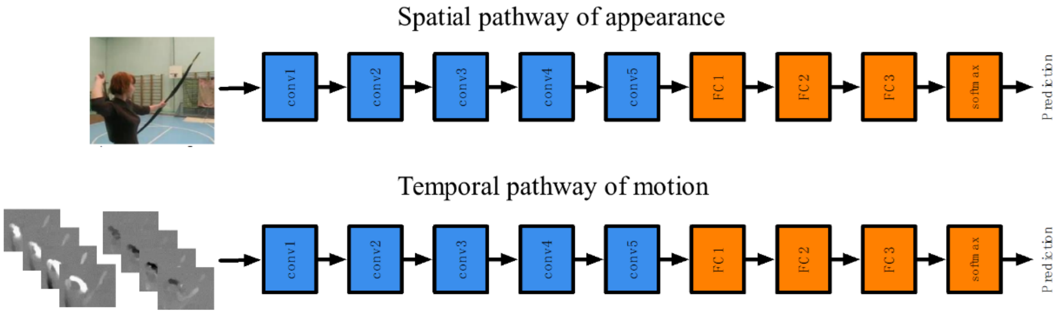

Figure 6.1: The architectures of the spatial CNN of appearance (top) and the temporal CNN of motion (bottom) using the VGG16 networks. Each pathway contains five convolutional layers and three fully-connected layers.

If we separately train the spatial CNN with the static images in action sequences and the temporal CNN with optical flow fields extracted from action videos, they will each yield independent recognition decisions, based solely on appearance or motion information respectively. To yield overall recognition of the action category given a video clip, the network architecture is further extended to achieve fusion of the multimodal information. The basic approach is to take the outputs of one layer in the spatial CNN and the outputs of one layer in the temporal CNN and treat them as the joint inputs to an additional network (usually a few FC layers) that perform holistic action recognition. While any layers may be selected for the information fusion, empirical studies conducted by Feichtenhofer, Pinz, & Zisserman (2016) suggest that the fusion of the outputs of the final convolutional layers (i.e., “conv5”) of both streams consistently yields the highest recognition accuracy across different datasets (see Figure

6.2). Accordingly, we adopted this fusion architecture for our two-stream CNNs model in simulations reported in the present paper.

More specifically, the two-stream CNN model uses the fusion layer to first stack the outputs from the “conv5” layers (each one is a 7 × 7 × 512 tensor) resulting in a 7 × 7 × 1024 tensor, followed by passing this stacked tensor to a convolutional layer consisted of 512 filters with a size of 1 × 1. The resulting output is still a 7 × 7 × 512 tensor. We follow a two-phase protocol to train the network: We first train the single stream networks shown in Figure 6.1 (spatial CNN and temporal CNN independently); We then integrate the convolutional layers of these two CNNs into the two-stream CNN (Figure 6.2) by fixing the parameters of the convolutional layers and only training the parameters of the FC layers.

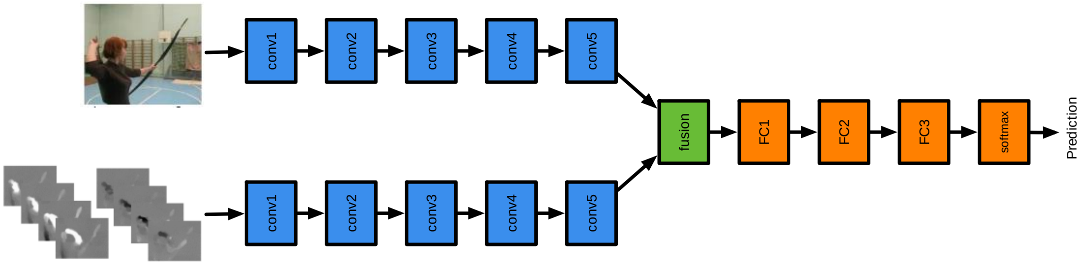

Figure 6.2: The architecture of the two-stream CNN. After training the spatial and temporal CNNs independently, the convolutional 5 layers of the two CNNs were fed into the fusion pathway which has 3 fully connected layers.

6.3.2 Stimuli and Procedure

Human motion capture data were selected from Human 3.6M dataset (Ionescu, Li, & Sminchisescu, 2011; Ionescu, Papava, Olaru, & Sminchisescu, 2014). The Human 3.6M dataset includes 47772 videos recorded in a lab setting through a motion capture system, with the same background and fixed camera locations. This dataset provides both raw videos with natural images (i.e. RGB videos) and also whole-body joint coordinates to generate skeletal and point-light displays of the actions from different viewpoints. A total of 15 categories of actions were included: giving directions, discussing something with someone, eating, greeting someone, phoning, posing, purchasing (i.e. hauling up), sitting, sitting down, smoking, taking photos, waiting, walking, walking dog, and walking together. Each action was performed by 7 actors. Each actor performed the same action twice, resulting in two recordings for each action.

RGB video training instances were generated by selecting 5s video clips with a non-overlapping 5s sliding window from the beginning of each recording to the end. Each video clip contains 150 frames with a 30 fps sampling rate. For each video clip, 5 additional versions were generated to increase the flexibility of the training data: (1) a zoom-in version with 200 pixels being cut-off from all the four boundaries; (2-5) spatially shifted versions with the human figure being shifted toward the top-left, top-right, bottom-left, and bottom-right corners which was achieved by cutting off 200 pixels from the corresponding boundaries. A total of 47772 video clips were generated in the end.

Because the Human 3.6M dataset provides motion capture data, in addition to the raw recorded RGB videos, we generated the skeletal and point-light displays using the tracked joint positions in actions. Skeletal and point-light videos were generated using the Matlab BioMotion toolbox (van Boxtel & Lu, 2013). Similarly, 5s video clips were selected from the long action recording sequences for all 15 categories. For skeletal displays, 7 different viewpoints were selected for each video clip, ranging from 30° counter-clockwise from the central viewpoint to 30° clockwise from the central viewpoint with a step size of 10°. A total of 13769 skeletal training instances were generated. Point-light videos were generated for the testing purpose. Only the central viewpoint was selected and a total of 1967 testing instances were generated. Sample frames of videos are shown in figure 6.3.

The training stage contains two steps. First, the model was trained with RGB videos separately for the spatial and temporal pathways, during which 90% of RGB training instances were used for training and the rest 10% were used as validation to test the model performance. After reaching a saturated accuracy, the trained weights of two pathways are saved. Second, a transfer learning was conducted by using the trained weights of RGB videos as initial values and retraining the entire CNN model with skeletal videos. Again, 90% of RGB training instances were used for training and the rest 10% were used as validation. In the testing stage, point-light displays were tested and the accuracy and confusion matrices were calculated.

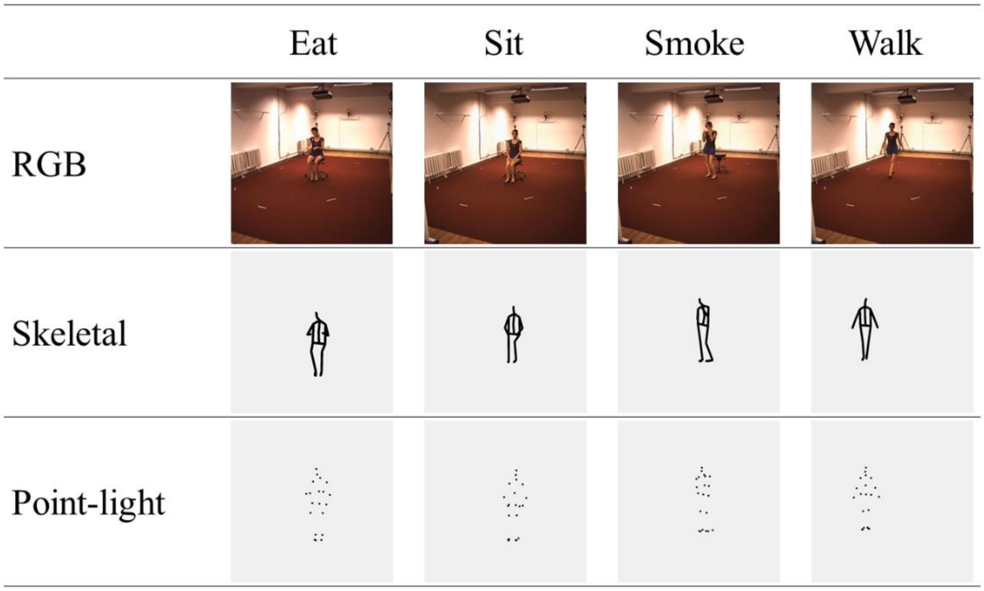

Figure 6.3: Sample training data in the RGB display and the skeletal display and testing data in the point-light display. Videos were generated from motion capture data of the Human 3.6M.

6.3.3 Results and Discussion

Training and testing accuracy is shown in Table 6.1. Transfer learning of skeletal displays reached a high accuracy in both pathways. For testing, the temporal pathway of motion yielded higher than chance-level accuracy (i.e. 0.56 with a chance level of 0.07) for action classification with point-light displays. In contrast, the spatial pathway of appearance showed poor generalization performance, yielding close to chance-level recognition accuracy (i.e. 0.16). These results support a crucial role for the motion processing pathway in action recognition for point-light displays. As demonstrated by confusion matrices as shown in Fig. 4, the spatial pathway for appearance was highly biased to certain action categories (e.g. discussing, making phone calls, and waiting) and did not show an ability to discriminate among other actions, whereas the temporal pathway classified all 15 actions with much better discrimination. The recognition errors made in the spatial pathway was likely due to the appearance similarities between static image frames of certain action categories. For example, sitting down was frequently judged as purchasing (i.e. hauling up) probably due to the similarity between body postures of these two action categories, with a signature of the posture of squat.

Even though the model was intensively trained on skeletal displays, the spatial CNN and the fusion network were not able to generalize from skeletal action stimuli to point-light action stimuli. The failure of the spatial CNN may not be surprising since actions in point-light displays reveal limited and sparse information about human forms. Interestingly, the test performance from the fusion layer (i.e. 0.27) was even lower than the average of two CNN pathways. From the confusion matrix (Figure 6.4, right), the fusion network showed similar results as the spatial CNN, suggesting that the fusion network gave much higher weights to the spatial CNN from the training with skeletal displays and failed to utilize the useful information passed on from the temporal CNN to perform well with the point-light actions. In other words, the fusion module was trained to integrate both appearance and motion information from the skeletal display, but did not establish optimal weighting between the two pathways for recognition of actions in the point-light displays.

Table 6.1. Training and testing accuracy in simulation1.

	RGB training	Skeletal transfer learning	Point-light testing
Appearance (spatial)	0.95	0.97	0.16
Motion (temporal)	0.79	0.99	0.59
Fusion (spatial + temporal)		0.99	0.27

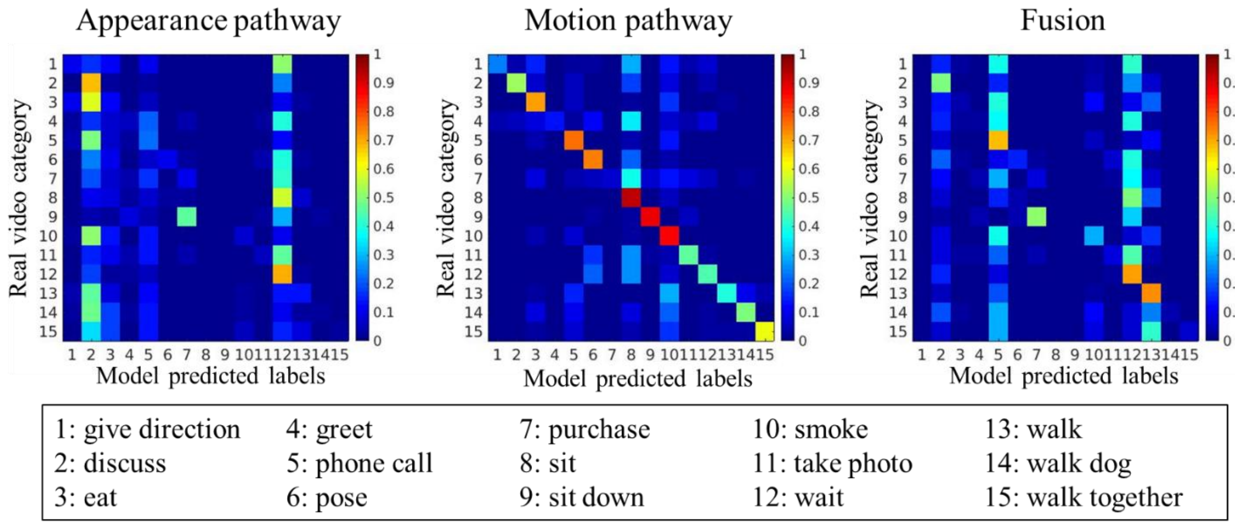

Figure 6.4. Confusion matrices of recognizing Human 3.6M point-light videos. From left to right, the matrices correspond to appearance pathway, motion pathway, and fusion respectively. Colors represent the proportion of videos being classified as certain categories. Proportions increase from blue to red.

6.4 Simulation 2

Results in simulation 1 showed that the two-stream CNN is able to differentiate actions after training with skeletal displays. In simulation 2, we further examined the performance of the two-stream CNN model on perceiving human actions with a violated causal relation between limb movements and body motions. The model is trained with intact actions with consistent facing and walking directions, and was tested with other action categories with inconsistent facing and moving directions, including moonwalk actions, backward actions, and in-place actions. If the CNN model is able to form a basic understanding of the implicit causal structure, the inconsistency between motion cues in three testing action conditions should introduce some confusion when judging the directions.

6.4.1 Stimuli and Procedure

Ninety-eight typical walking actions were selected from the CMU motion capture database. Skeletal walkers were presented in either a left profile view or a right profile view, yielding 98 instances facing left and 98 facing right. Four conditions of actions were generated: (1) intact action, (2) moonwalk action, (3) backward action, and (4) in-place walking action. First, intact actions are the typical walking actions with consistent limb movements and body motions, and also consistent facing direction and walking direction. Second, moonwalk actions were generated by reversing the horizontal moving direction of the global body translation while keeping the limb movements sequence intact. In other words, when a walker moves limbs in a way to naturally propel her body moving left, the body, however, moves to the right in the moonwalk actions. This is analogues to the well-known example of Michael Jackson's moonwalk dance movement. Third, backward walking actions were generated by reversing the entire motion sequence, so limb movements and body motions still share a consistent causal relation but the walking direction is opposite to the facing direction. Finally, in-place actions were generated by removing the global body translation component but only keeping the limb movements, so the walker looks like walking on a treadmill. Categories of all videos were defined based on the "facing direction", meaning if the face is facing left or right regardless of limb or body movements. As shown in figure 6.5, in the causal condition, a "facing right" instance would show a walker facing right and walking towards the right. In the moonwalk condition, a "facing right" instance would show a walker facing right with limb movements indicating walking towards right but body translation moving towards left.

Intact actions were used as training stimuli in the transfer learning while the other three conditions were used as testing stimuli. All the training and testing motion stimuli were presented in the skeletal form. Transfer learning was conducted based on the trained weights of Human 3.6M skeletal videos. The original dense layer with a softmax activation with 15 categories was replaced by a new softmax dense layer with 2 categories (facing left vs. facing right walking). Only the last layer of the model was retrained with a goal of distinguishing the walking directions with weights of all previous layers being fixed. The retrained model was then tested on the other three conditions. Accuracy and confusion matrices were calculated for each condition.

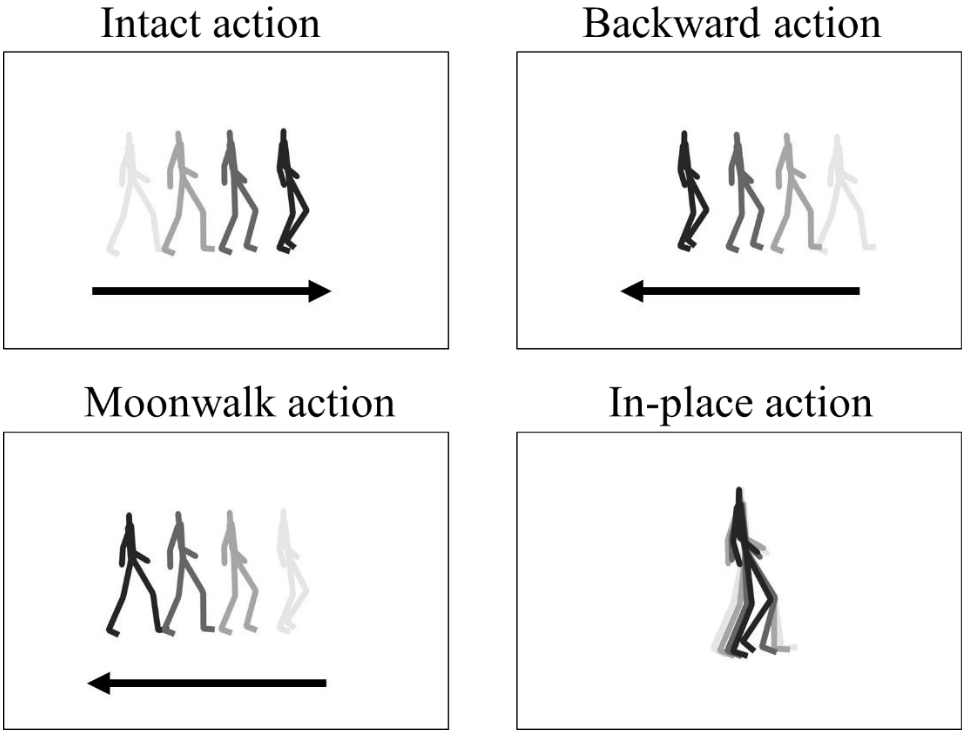

Figure 6.5. Illustrations of the four walking action categories: intact action, moonwalk action, backward action, and in-place action. Here, illustrations of all conditions show instances of the "facing right" category. Each plot shows several possible limb movements for a stick-figure resulting from posture changes over time. The sticks in the walker change from light to dark color to denote elapsed time. Arrows below all the stick-figures, except the in-place action, indicate the global body translation direction. In the in-place condition, the walker remained in a stationary location.

6.4.2 Results and Discussion

Transfer learning of intact actions reached high accuracy (i.e. 100%) for both the spatial and temporal pathways. Testing results are shown in table 6.2. The model trained with intact action judged other categories consistently based on the facing direction while ignoring the body translation component. For example, the model reached an accuracy of 0.95 for the moonwalk condition in the spatial pathway, meaning that for 95% of all moonwalk instances, the video “correctly” classified videos based on the facing direction, even though the body translation is totally inconsistent with facing direction. This suggests that the spatial pathway of appearance was highly biased by body posture information and the global body translation is not determining the classification results. Results are similar for the backward condition (i.e. 100%). Interestingly, for in-place actions, the results (i.e. 84%) were not as decisive as moonwalk or backward conditions. This is probably due to an appearance difference between causal and in-place actions. For example, the spatial location of the walker in a causal condition varies across time, but the walker in the in-place condition is always at the center. This mismatch of appearance may have confused the model trained with intact actions. The results are not surprising since the spatial pathway does not have the temporal information of motions, so the judgment can only be based on static posture appearance.

The temporal pathway showed slightly different results compared to the spatial pathway. Moonwalk instances were judged inconsistently based on either facing direction or the body translation direction, yielding a low score of 0.69. This result suggests that the temporal pathway of motion may take into consideration of both the body translation component and the limb posture component when classifying actions. However, the overall judgment is heavily biased toward the facing direction compared to the body translation component. The results of backward actions were similar to moonwalks, again suggesting that both motion components were considered when the model was making judgments. Even though moonwalk actions and backward actions are different in terms of the causal relation between limb movements and body translation, the current model was not able to reflect that difference in terms of classification accuracy. The accuracy score drops from 0.84 in the spatial pathway to 0.79 in the temporal pathway for the in-place actions. This may be due to that the temporal pathway gives higher weights to the body translation motion cues when making decisions but in-place actions lack an informative indicator of body translation direction, which hinders making correct judgments of the facing direction.

The performance of fusion fell between the two pathways, indicating a compromised decision based on inputs from two pathways. The current result may be caused by a missing representation of the global temporal pattern of videos since the entire video clip was cut into chunks of 10 frames (0.33s) during the training. This may significantly limit the representation of how the body moves across space in a longer range of time. Thus, it may not be surprising that higher weights were given to the static body posture instead of body translations.

Table 6.2. Training and testing accuracy with confusion matrices in simulation 2. From top to bottom, the rows show results of the spatial pathway, temporal pathway, and fusion. Confusion matrices show the proportion of model predictions. The first and second row indicate videos with a ground truth of facing left and facing right respectively. The first and second column indicate videos being judged as facing left and facing right respectively.

	Intact action transfer learning	Moonwalk action testing	Backward action testing	In-place action testing
Spatial (Appearance)	score: 1.0 [1 0 0 1]	score: 0.95 [0.96 0.04 0.06 0.94]	score: 1.0 [1 0 0 1]	score: 0.84 [0.71 0.29 0.02 0.98]
Temporal (Motion)	score: 1.0 [1 0 0 1]	score: 0.69 [0.77 0.23 0.38 0.62]	score: 0.68 [0.70 0.30 0.34 0.66]	score: 0.79 [0.93 0.07 0.36 0.64]
Fusion	score: 0.975 [0.99 0.01 0.04 0.96]	score: 0.866 [0.94 0.06 [0.21 0.79]	score: 0.90 [0.95 0.05 0.16 0.84]	score: 0.78 [0.78 0.23 0.22 0.78]

6.5 Simulation 3

Simulation 2 showed that the two-stream CNN model was not able to dintinguish moonwalk from backward actions, although the former violates the causal relation between limb movements and body motion and the latter has the causal congruency. This failure may be caused by a missing of the global temporal structure of videos.

Besides CNNs, another type of deep neural network is the recurrent neural network (RNN), which is widely used for processing sequential input data, such as handwriting recognition (e.g. Graves et al., 2009; Sutskever, Martens, & Hinton, 2011) and speech recognition (e.g. Graves, Mohamed, & Hinton, 2013). Different from feedforward CNNs which only consider the current input, RNNs maps input sequences to a sequence of hidden states, and then provide outputs based on the sequential data. Each hidden states considers not only the current input but also integrates information from previous states. This enables the learning of complex temporal dynamics. In action recognition, The combination of CNN and RNN enables us to not only use the CNN’s power to learn compact representations of image data, but also employ the RNN’s power to learn temporal patterns of the entire video which is necessary for human pose prediction.

Specifically, long short-term memory (LSTM) (Hochreiter, & Schmidhuber, 1997) is a type of RNN architecture which was proposed as a solution for the vanishing gradient problems of RNNs, referring to the problem that when a model has many layers, the gradient decreases exponentially as it propagates down to the initial layers and prevents efficient training procedures. LSTM module contains many recurrently connected memory blocks, enabling integrating and forgetting information over time. This mechanism mirrors the mechanism of working memory in human brains, that dopamine is released to enable forgetting and the formation of new memories with plasticity (Berry, Cervantes-Sandoval, Nicholas, & Davis, 2012). Many recent works have witnessed the success of LSTM in action recognition (Liu, Shahroudy, Xu, & Wang, 2016; Donahue et al., 2015) and pose prediction (Fragkiadaki, Levine, Felsen, & Malik, 2015; Martinez, Black, & Romero, 2017).

In simulation 3, we added an LSTM model on top of the two-stream CNN. We aim to test if the added LSTM module will enable the model to have a representation of the global temporal pattern of actions, and if the model is able to represent moonwalk and backward actions differently.

6.5.1 Model structure

As shown in figure 6.6, an LSTM component was added to the FC2 layer of CNN and replaced the original softmax layer. Each time point corresponds to one hidden state of LSTM and all the hidden states were connected across time. The LSTM module takes features extracted from the FC2 layer as inputs. For the image pathway, every 10 static frames were treated as a unit and features of only the middle frame were extract to enter the LSTM module. For the flow pathway, every 10 consecutive frames of optical-flow images were grouped as a unit and one set of FC2 feature was extracted based on the entire 10 frames. In the end, each 5s video (i.e. 150 frames in the raw video) yield 15 time steps of FC 2 features. Each time step enters the LSTM module consecutively and the LSTM module provides one output at the end of actions.

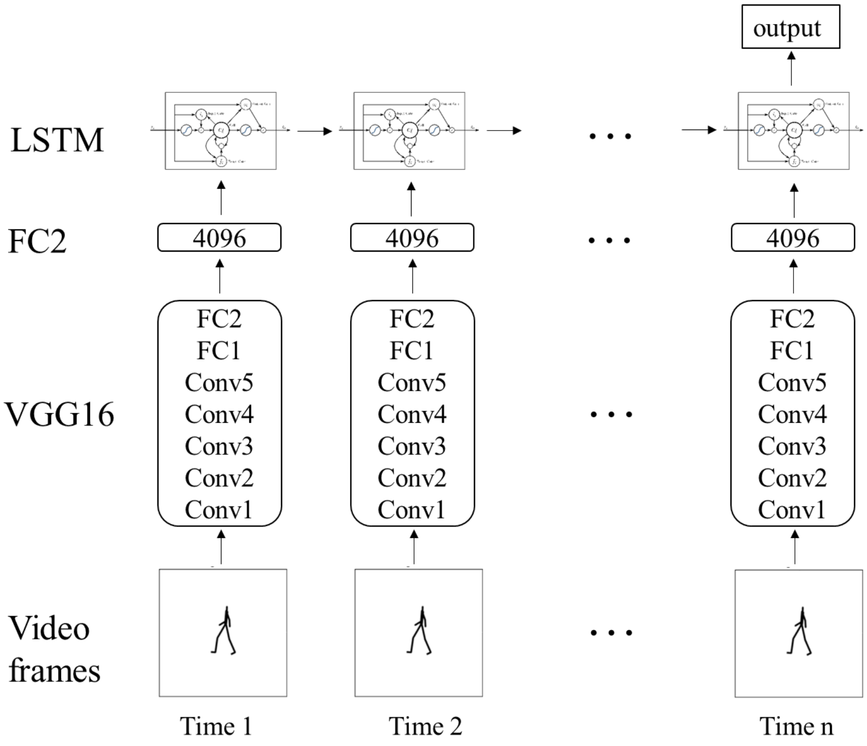

Figure 6.6. An illustration of the model structure after combining CNN and LSTM modules. One LSTM cell is connected to the feature outputs of FC2 at each time step. LSTM cells are responsible for keeping track of the dependencies between the elements in the input sequence. Each LSTM cell contains an input gate, an output gate, and a forget gate. The input gate controls the extent to which a new value flows into the cell. The forget gate controls the extent to which a value remains in the cell and the output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit.

6.5.2 Stimuli and Procedure

Same video stimuli were used in simulation3 as in simulation 2. To train the new model with an additional LSTM module, skeletal display of Human 3.6M dataset were used as training instances. Weights of the CNN model were fixed and only the weights associated to the LSTM module were updated. After reaching a saturated accuracy, the model weights were saved for the purpose of transfer learning. Again, intact action video clips were used as training instances to retrain the last softmax FC layer connected to LSTM, with the goal of differentiating left facing actions from right facing actions. Then, the other three action conditions were used as testing instances to calculate the accuracy and confusion matrices.

To test if the model can differentiate intact actions from moonwalk actions, a transfer learning was further conducted to retrain the softmax FC layer to distinguish intact actions from moonwalk actions, regardless of the facing direction. Accuracy and confusion matrices were calculated. The same procedure was also conducted for the transfer learning of distinguishing causal from backward actions. To compare the representation of videos, features were extracted from the LSTM layer for each video. The output contains the LSTM representation of the entire video, containing 3000 nodes. Thus, each video gets a 1 × 3000 vector LSTM representation. Representational dissimilarity matrix (RDM) was calculated based on the cosine distances between feature vectors of pairs of videos.

6.5.3 Results and Discussion

The accuracy of transfer learning and testing are shown in table 6.3. After adding the LSTM component, the testing classification score of facing directions of walkers dropped to close to zero for both moonwalk actions and backward actions, as the model discriminate walking directions in accordance to the global body displacement which is opposite to the facing direction. This suggests that adding LSTM in addition to CNN enabled the learning of temporal patterns of the entire video, which made the direction of body translation an important motion cue in the classification of facing directions. However, this trade-off between weighting limb postures and body translations also seems to happen in a winner-take-all manner. In contrast to human observers who would consider both how limbs move and how the body translates and the relation between the two motions, the current model with LSTM module seems to be highly biased by the direction of body translation. This is further supported by the performance of in-place actions (i.e. 61% for the spatial CNN and 60% for the temporal CNN), which is close to the chance level, suggesting that when the body translation information is not available, the model fails to distinguish facing directions of in-place actions.

To examine if the model can distinguish causal from moonwalk actions, a transfer learning was conducted to retrain the softmax layer to distinguish the two categories (i.e. actions with causally congruent motion cues or actions with incongruent motion cues). Results are shown in table 6.4. The appearance pathway yields an accuracy of 0.59, which is not greatly different from the chance level 50%. This suggests that only for 59% of time, the model was able to distinguish causal from moonwalk actions. This failure of transfer learning suggests that the model was not able to learn the relation between limb movements and body translation, regardless of being able to capture the global temporal pattern of the videos. The motion pathway did better on the transfer learning, yielding an accuracy of 83%, suggesting that with optical flow information of limb movements combined with global temporal patterns, the model could better distinguish causal from moonwalk actions with an accuracy of 83%. However, this is still much lower compared to the previous transfer learning results on discriminating left vs. right facing directions, indicating that the task of discriminating causal vs. moonwalk is more difficult to the model compared to a task of discriminating facing directions.

The transfer learning between intact actions vs. backward actions yielded slightly better performance in the spatial pathway (i.e. a score of 0.87) than the previous transfer learning between causal and moonwalk actions (i.e. a score of 0.59). This increased performance may be due to a difference in the similarity to intact actions, since moonwalk actions share the same limb movement posture sequences as intact actions across time, while backward actions share entirely opposite limb movement sequences compared to intact actions. The temporal pathway of the two transfer learning yielded similar results (i.e. 0.83 for moonwalk actions and 0.85 for backward actions), indicating that the model was not able to distinguish the causal congruent or causal incongruent relations between limb movements and body motions.

To rule out the possibility that the low performance in the previous transfer learning is due to lack of learning in the temporal integration module of LSTM, we further retrained the LSTM layer in addition to the softmax FC layer on the task of discriminating causal vs. moonwalk actions, as well as causal vs. backward actions. With more adjustments on parameters via training, the new transfer learning yield higher performances (e.g. > 90% for all the pathways). To compare the representations of different video categories, RDMs were calculated by getting cosine distances among LSTM representations between all pairs of videos from the four categories. For the visualization purpose, the cosine distance values were averaged within the comparison of one category to another (e.g. causal facing left vs. moonwalk facing left). Representations of the transfer learning of intact walking vs. backward walking yielded highly similar RDMs as intact walking vs. moonwalk so only results of the latter were shown in Fig 7. Both appearance and spatial pathways showed no clear distinction between intact walking, moonwalk, or backward walking actions. Only the in-place actions yielded a clear distinction from all the three other action categories. This result suggests that even after a transfer learning with clear goals of distinguishing causal from moonwalk actions, the representations still didn’t reveal a clear distinction.

Table 6.3. Training and testing accuracy with confusion matrices in simulation 3. From top to bottom, the rows show results of the spatial pathway and the temporal pathway, which served as inputs to the LSTM module. Confusion matrices show the proportion of model predictions. The first and second row indicate videos with a ground truth of facing left and facing right respectively. The first and second column indicate videos being judged as facing left and facing right respectively.

	Intact action transfer learning	Moonwalk action testing	Backward action testing	In-place action testing
Appearance (spatial)	score: 1.0 [1 0 0 1]	score: 0.03 [0.04 0.96 0.99 0.01]	score: 0.02 [0.04 0.96 1 0]	score: 0.61 [0.91 0.09 0.69 0.31]
Motion (temporal)	score: 0.99 [1 0 0.02 0.98]	score: 0.19 [0.38 0.62 1 0]	score: 0.07 [0.08 0.92 0.94 0.06]	score: 0.60 [0.41 0.59 0.21 0.79]

Table 6.4. Transfer learning performance and confusion matrices of distinguishing intact action from moonwalk or backward actions. The first two columns show the results of only retraining the softmax FC layer. The second two columns show the results of retraining both the LSTM module and the softmax FC layer.

	Transfer learning on the softmax layer		Transfer learning on the LSTM + softmax layer	
	Causal vs. Moonwalk	Causal vs. Backward	Causal vs. Moonwalk	Causal vs. Backward
Appearance (spatial)	score: 0.59 [0.51 0.49 0.32 0.68]	score: 0.87 [0.89 0.11 0.15 0.85]	score: 0.99 [0.99 0.01 0.01 0.99]	score: 0.94 [0.95 0.05 0.07 0.93]
Motion (temporal)	score: 0.83 [0.82 0.19 0.15 0.85]	score: 0.85 [0.83 0.17 0.13 0.87]	score: 0.99 [0.98 0.02 0.01 0.99]	score: 0.92 [0.89 0.11 0.06 0.94]

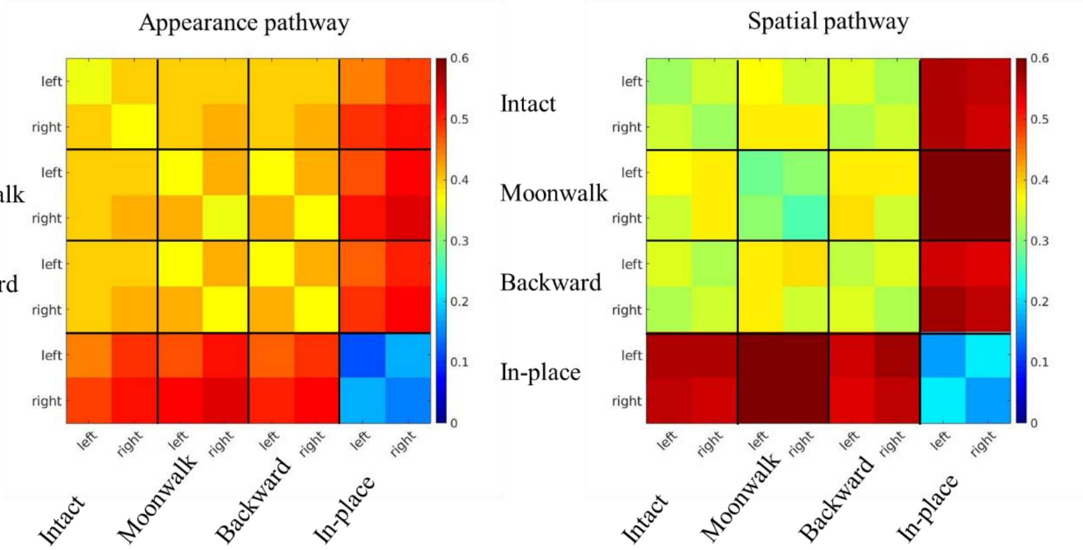

Figure 6.7. Representational dissimilarity matrices (RDMs) of the four action categories calculated from the LSTM features extracted after transfer learning of intact actions and

moonwalk actions. The two plots show the RDM of the appearance pathway and the spatial pathway respectively.

6.6 General Discussion

Through three simulations, we assessed whether a two-stream CNN for action recognition can be generalized across display forms and if the CNN has an understanding of the causal relation between motion cues. Simulation 1 showed that despite the high accuracy after training on skeletal displays, the model performed poorly when being tested on the novel point-light displays. The two-stream CNN model trained with natural videos were based largely on appearance similarities between static image frames, such as colors, textures, scales of objects, and background scenes. The model appears to heavily rely on statistic regularities in the appearance and optical flow fields of videos, rather than forming robust representations of human body movements per se. Hence, two-stream CNNs proved susceptible to a mismatch between training and testing datasets, reflecting less robust representations of human body movements.

Previous work has provided mounting evidence that convolutional neural networks employ different representations of objects than does the human visual system. Studies based on "adversarial examples" (Szegedy et al., 2014) or "fooling examples" (Nguyen, Yosinski, & Clune, 2015) clearly showed that CNNs can be easily fooled by images altered with minimally perturbations that are imperceptible to human observers. Furthermore, Geirhos et al., (2017) compared CNNs with human observers on a series of object recognition tasks and concluded that the human visual system is much more robust to a range of image manipulations such as contrast reduction, adding noise or novel eidolon-type distortions than CNNs. Similar results were reported by Dodge & Karam (2017). Using highly distorted images, these researchers found that humans greatly outperformed neural networks trained on clean images even when the images were presented for only 100 ms. Although CNNs can match or even outperform humans on some recognition tasks, there are marked differences in how humans and CNNs process visual information.

Simulation 2 showed that the two-stream CNN was not able to discriminate the difference between actions based on the consistency between motion cues in actions. In moonwalk actions, even though the limb movements show great inconsistency with body translation, the two-stream CNN model was not sensitive to this violation of motions. This was likely due to a lack of representation of the global temporal pattern of the video, which made the model fail to recognize global body translations as an important motion cue. In simulation 3, an LSTM module was added on top of CNN to enable the representation of global body displacement patterns of videos. Results showed that LSTM indeed showed sensitivity to global body translations. However, this did not help the model to fully understand the causal relation between limb movements and body translations.

Across simulations, the temporal pathway of motion yielded better performance than did the spatial pathway when testing actions presented in skeleton displays or point-light displays. This result suggests that the representation based on motion information is more robust and less susceptible to changes in the videos, such as scales and body structures. In addition to increasing the model depth or expanding the training dataset, we suspect that future advances based on CNNs will require robust weighting strategies for the integration of the two pathways. In addition, using controlled stimuli commonly used in psychophysics provides a useful test set to assess the generalization ability of CNNs, and to gauge the underlying representational commonalities and differences relative to the human visual system.

Together, the evidence shows a lack of causal representation of human actions as in the current form of deep neural networks. While neural networks are efficient in learning the statistical regularities in stimuli, they lack explicit representations of the high-level meaning and relations that cannot be directly perceived from visual stimuli. The totally stimulus-driven approach also leads to the requirement of large numbers of training instances regardless of the task, which is highly distinct from the way of human learning new items or concepts through very few exemplars, even demonstrated by young babies (Carey & Bartlett, 1978; Landau, Smith, & Jones, 1988). Achieving human-like ability in action perception may require a system with not only stimulus-driven training procedures, but also a mentalizing system, or "theory of mind" (ToM) (Van Overwalle, & Baetens, 2009) with potential hierarchy structure to infer the cause and effect at different stages of observations. Similar ideas have been demonstrated in one-shot learning of written digits and image classification (Lake, Salakhutdinov, & Tenenbaum, 2015; Ren et al., 2018) but future studies should be done in the field of human action recognition since the understanding of the causality between motion cues is the foundation of the understanding of goal-oriented human actions.

CHAPTER 7

General Discussions and Conclusions

In summary, my dissertation has investigated (1) the role of causal constraints in human action perception, (2) the emergence of causal action perception in infants, (3) the neural mechanism underlying causal perception of actions, and (4) the performance of computational models on the recognition of causal actions.

Chapter 2 presented a behavioral study that shows that the causal constraint of human actions plays an important role in the process of perception and making inferences with body movements. Human observers judged actions with motion cues that follow a causal relation to be more natural. In addition, the causal expectation for human body movements not only affects perceptual impressions regarding the naturalness of observed actions but also guided the interpretation of motion cues within a more generalized causal context. Importantly, this work pinpointed the role of causality using manipulations based on the temporal-priority principle and helps to dissolve the ambiguity between a truly causal relation and an inconsistency between motion cues. Chapter 3 further presented behavioral evidence of a top-down role of causality in facilitating a continuous perception of human actions. When there is a causal connection between motions of agents and physical objects, actions are more likely to be perceived as continuous. Both chapters showed that causality serves as a bridge that connects action perception and action understanding. On the one hand, the perceptual process integrates low-level motion information and forms a causal representation, which enables the understanding of goals of actions. On the other hand, the prior knowledge and expectations of an understanding of action influence action perception in a top-down manner and facilitate the formation of a robust and meaningful perception.

Chapter 4 presented developmental evidence showing that by 18 months, infants are sensitive to the motion consistency constraint that governs human actions. In addition, the development of causal action perception may be in parallel with the development of motor functions. The results show a close connection between action perception, action understanding, and even action performance. The early emergence of causal understanding may root from observing co-occurrence of motion cues in the environment and also from knowledge gained from experiencing outcomes of actions through infants' interactions with others and the environment. Reciprocally, the knowledge of causal connections between actions and outcomes further reshapes the perceptual process and guides the development of gross motor skills.

Chapter 5 presented a MEG study which revealed the spatiotemporal neural activities underlying the visual processing of causal constraints in biological motion. This work extended previous works with simple physical causal events to human actions. By contrasting the neural activities of action perception and causal reasoning, the results showed that the causal representation occurs after the action recognition processing in time and involves a more distributed neural network including brain areas in the parietal, temporal, and frontal lobes that are responsible for spatial relation reasoning, decision making, and intention understanding. The findings highlight a distributed neural network that supports human action understanding and causal reasoning, involving cortical regions beyond visual areas that process low-level motion signals.

Finally, Chapter 6 presented computational simulation results of a two-stream CNN model with added LSTM module and showed the limitations of the state-of-art deep neural networks in accounting for causal perception in human actions. While the current neural networks emulate human visual pathways and performed well on tasks based on low-level visual information, they lack a representation of high-level reasoning processes that build the foundation of intention understanding and goal inference. The findings advocate for putting more attention in the communication between the development of computational models from an engineering aspect and the investigation of mechanisms of the human cognitive system. Even though the CNN models nowadays achieve human-like performance on some visual tasks, it will need to incorporate representations that cannot be directly extracted from low-level visual information (e.g. causality, intentions, semantic relations) to support reasoning and understanding of visual inputs in a more robust and adaptive manner.

Overall, the findings from behavioral, developmental, and neuroimaging studies showed converging evidence that human action processes go far beyond pattern recognition of motion signals. In addition, action understanding is supported by reasoning with relations between motion cues. The development of causal understanding is achieved not only through visual observations but also through causally interacting with the world. In contrast to findings from human experiments, simulations of deep neural networks revealed the limitations of the current modeling approach to account for human action processing. Most current neural networks are built in a pure stimulus-driven manner that excels in extracting statistic regularities from visual stimuli and recognizing patterns based on these learned statistic regularities. This modeling architecture is analogous to a human brain with only the bottom-up connections of visual areas located in the occipital cortical region, while missing critical modules from parietal, temporal and frontal cortical regions that process abstract relational properties and make decisions. The lack of learning mechanisms for high-level knowledge greatly limits the efficiency and generalizations of deep learning models. To build artificial intelligence that can achieve action understanding and meaningful social interaction, it would require a generative model that has a basic understanding of hierarchical structures of human body motions and causal relations between motion components.

In conclusion, the research in the thesis highlights that causality is a key factor in human action perception and it serves as a bridge between action perception and goal understanding. Although this work provides candidate answers to some of the important questions in the field of action perception, many questions remain to be answered. Future research might explore the links between causal perception and understanding from several aspects.

7.1 How Does Causality Binds Actions and Outcomes?

In previous chapters, it has been shown that causality is the cement of action perception and understanding. However, the mechanism that supports this function of causality in actions remains unknown. One hypothesis roots from the idea of temporal contiguity, which suggests that causality gives rise to the sense of agency by the consistent cause-effect pairings in the daily life (Wegner, 2004; Young, Rogers, & Beckmann, 2005). In other words, causality connects perception and understanding through time. The co-occurrence of having a conscious thought prior to an action, the actual action performance, and the action outcome, gives us the sense of agency, or conscious experience of the authorship to the performed action. For example, the perceived time elapse between an intentional action (e.g., a key press) and its subsequent sensory outcome (e.g., a tone or flash) is compressed, such that all sensory events following an action appear to draw closer in time to that action. This idea has been tested in the context of physical events, and causality has been shown to yield a temporal binding effect of causally connected events. For example, Haggard, Clark, Kalogeras, (2002) showed that voluntary actions and their consequences are shifted towards each other in subjective time. Several other studies found that factors that influence the perceived causality between the voluntary action and its consequence, such as temporal contingency and contiguity, also modulate the temporal binding between them (Shanks, Pearson, & Dickinson, 1989; Moore and Haggard, 2008; Moore et al. 2009). Buehner and Humphreys (2009) have shown that the causal relation between action and consequence is more important than intentionality for the phenomenon to occur. Cravo, Claessens, & Baldo (2009) further dissociated causality from voluntary actions and showed that causality and voluntary actions together determine the temporal binding effects in launching events.

However, studies so far have been limited to relatively simple physical events and remain to be generalized to natural scenarios with human actions. Future studies could investigate the impact of causality on the time perception of human actions. The perceived time duration of video clips that entail causal or non-causal human actions could be measured. Thus, if causality glues causes and effects in time, we would expect that the two causal video clips would be perceived to have a shorter duration and the non-causal video clip pairs would be perceived to have a longer duration.

7.2 How Does Infants Learn Causal Constraints in Human Actions?

As shown in previous literature and the developmental work in this thesis, the ability to perceive causality from dynamic events seem to emerge early in infancy (Leslie and Keeble, 1987; Cohen and Amsel, 1998, Desrochers, 1999). This ability may be the key to determine the success of social interactions and to generate predictions based on observed actions. For example, infants need to first have an understanding that people's actions are usually goal-oriented. Then, they can observe the actions of others and consider the intention to generate a prediction about future actions and prepare to respond accordingly in an appropriate way (Southgate, & Vernetti, 2014). Even though this work has shown an early emergence of adopting the causal constraints in the perception of human actions, it still remains unclear what important factors enables the learning of causal constraints at the early stage of life. Potential candidates include but not limited to (1) visual information such as actions of parents which provide the repeated exposure to the spatiotemporal congruent cause-effect binding, as well as gestures and facial expressions from other people's action; (2) motor experience such as the consistent action-outcome pairing; (3) the development of language, which may facilitate the abstract representation of causality; and (4) emotions and rewards, which may consolidate the learning process based on reinforcements. Future studies could investigate how these factors influence the development of causal perception and action understanding based on individual differences. For example, recent years have witnessed an increasing use of head-mounted video cameras as a method to study the visual environments of infants and young children (e.g. Smith, Yu, Yoshida, & Fausey, 2015). This enables researchers to capture the "child's view" and have a better understanding of the relation between visual inputs and behavior outcomes.

In addition, it would be important to examine the development of causal understanding in abnormal populations such as children with autism spectrum disorders (ASD). In children with ASD, difficulties in understanding and predicting others' actions, as well as difficulties in theory of mind and social-cognitive skills, are frequently documented (e.g. Cattaneo et al., 2007; Zalla, Labruyère, Clément, & Georgieff, 2010). With the advancement of research on ASD, symptoms in the areas of intentional communication, motor development and social and emotional development can be detected as early as in the period of 12 – 24 months in infants (e.g. Nadel, &

Poss, 2007). Given the importance of causality in action understanding, it may worth investigating the development of causal perception in children who are diagnosed with early symptoms of ASD. One previous study showed that children with ASD were also able to show an adaptation aftereffects for perceptual causality as the control group (Karaminis et al., 2015). However, the effect remains to be tested in the context of human actions and social interactions.

7.3 Can Causal Binding be Revealed by Neural Oscillation?

The MEG study in the thesis provides evidence on the temporal and spatial representations in brain to support causal perception in human actions. However, the neural mechanism of how causality facilitates the binding of motion cues remains unclear. One possible neural mechanism of binding problems is stimulus-locked neural oscillation. The hypothesis is that neural oscillations can code role bindings by having pools of neurons respectively representing a role and an object fire in synchrony if they are bound together and out of synchrony if not (von der Malsburg & Buhmann, 1992). Empirical evidence from EEG studies also suggests that phase synchrony within the frontoparietal network can bind object properties together in working memory (Phillips, Takeda & Singh, 2012). In addition, several computational models of relational reasoning and cognitive control have employed neural oscillations as a basic mechanism enabling working memory to code the bindings of objects into relational roles (Doumas, Hummel & Sandhofer, 2008; Hummel & Holyoak, 1997, 2003; Shastri & Ajjanagadde, 1993). Recent years have witnessed the use of behavioral oscillation paradigm as a method to tap into oscillatory phenomena based on psychophysical measurements of response time and accuracy (for a review see VanRullen & Dubois, 2011). For example, Lu, Morrison, Hummel and Holyoak (2006) showed that visual information presented as temporal flicker in the gamma band can modulate the perception of spatial relations between multiple objects in a subsequent display. Also, Peng, Javangula, Lu and Holyoak (2018) showed that theta- and alpha-band oscillations in reaction time were evoked by a task that required relational role binding. In addition, the authors also found a phase-shift was observed between the two semantic roles.

Based on previous findings, it may worth investigating if the representation of different motion components in human actions that form causal relations is achieved by synchronous activation patterns of neurons. The behavioral oscillation paradigm as well as brain stimulation methods, such as transcranial alternating current stimulation (tACS), can open a window for us to examine how causality is encoded in the brain and how different components in a causal relation are bound in time.

APPENDICES

APPENDIX A: MOTOR QUESTIONNAIRE USED IN CHAPTER 4

Gross motor skills

In the following please think about your child's gross motor skills and motor control. These skills relate to how easily your child is able to control his or her own movements, orient, obtain toys, or move around the room.

Sure that child does NOT show behavior (e.g., you have seen your child fail when attempting this or a similar behavior)	Child probably does NOT show behavior yet	Unsure whether child could do this or not (please try to use this category seldom)	Child probably shows this behavior	Sure that child shows this behavior and remember a particular instance
-2	-1	0	1	2

When rating a behavior your child used to do but that is not developmentally appropriate anymore (e.g., crawling when the child is already walking alone) please rate this behavior as +2.

1. When sitting on your lap with back support provided by you, your child will hold his/her head up and steady when looking around the room?	-2	-1	0	1	2
2. When placed into a crawling position resting on hands and knees, your child will crawl forward for a few steps (3-5)?	-2	-1	0	1	2
3. When placed into a sitting position on the floor, your child is able to hold on to some furniture and pull into a standing position	-2	-1	0	1	2
4. When placed into a standing position, your child will stand alone for a few seconds without help	-2	-1	0	1	2
5. When moving around freely, your child will run short distances around the room	-2	-1	0	1	2
6. jump down from boxes, small steps, or similar without falling	-2	-1	0	1	2
7. When walking down a hallway or small room, your child will walk in a straight line with arms lowered and swinging freely	-2	-1	0	1	2
8. During free play or pretend play, you notice your child is able to walk backwards for several (5 or more) steps?	-2	-1	0	1	2

9. During free play or pretend play, you notice your child is able to jump forward over small obstacles such as a curb or box?	**-2**	**-1**	**0**	**1**	**2**
10. When placed in front of a flight of stairs, your child is able to walk up stairs without aid? (4 or more steps)	**-2**	**-1**	**0**	**1**	**2**

References

Abravanel, E., Levan-Goldschmidt, E., & Stevenson, M. B. (1976). Action imitation: The early phase of infancy. *Child Development, 47,* 1032–1044.

Ackerman, C. M., & Courtney, S. M. (2012). Spatial relations and spatial locations are dissociated within prefrontal and parietal cortex. *Journal of Neurophysiology*, *108*(9), 2419-2429.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in cognitive sciences*, *4*(7), 267-278.

Assmus, A., Marshall, J. C., Ritzl, A., Noth, J., Zilles, K., & Fink, G. R. (2003). Left inferior parietal cortex integrates time and space during collision judgments. *Neuroimage*, *20*, S82-S88.

Baker, C. L., & Cynader, M. S. (1986). Spatial receptive-field properties of direction-selective neurons in cat striate cortex. *Journal of Neurophysiology*, *55*(6), 1136–1152.

Baker, Curtis L, & Braddick, O. J. (1985a). Temporal Properties of the Short-Range Process in Apparent Motion. *Perception*, *14*(2), 181–192.

Baker, Curtis L., & Braddick, O. J. (1985b). Eccentricity-dependent scaling of the limits for short-range apparent motion perception. *Vision Research*, *25*(6), 803–812.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child development*, *72*(3), 708-717.

Bardi, L., Regolin, L., & Simion, F. (2011). Biological motion preference in humans at birth: Role of dynamic and configural properties. *Developmental Science*, *14*(2), 353-359.

Bates, E., Carlson-Luden, V., & Bretherton, I. (1980). Perceptual aspects of tool using in infancy. *Infant Behavior and Development*, *3*, 127-140.

Bechlivanidis, C., & Lagnado, D. A. (2013). Does the “why” tell us the “when”? *Psychological Science*, *24*(8), 1563-1572.

Bechlivanidis, C., & Lagnado, D. A. (2016). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58–66.

Berry, J. A., Cervantes-Sandoval, I., Nicholas, E. P., & Davis, R. L. (2012). Dopamine is required for learning and forgetting in Drosophila. *Neuron*, *74*(3), 530-542.

Bertenthal, B. I., Proffitt, D. R., & Kramer, S. J. (1987). Perception of biomechanical motions by infants: implementation of various processing constraints. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(4), 577.

Bidet-Ildei, C., Kitromilides, E., Orliaguet, J. P., Pavlova, M., & Gentaz, E. (2013). Preference for point-light human biological motion in newborns: contribution of translational displacement. *Developmental Psychology*, *50*(1), 113.

Blos, J., Chatterjee, A., Kircher, T., & Straube, B. (2012). Neural correlates of causality judgment in physical and social context—the reversed effects of space and time. *NeuroImage*, *63*(2), 882-893.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, *16*(11), 3737-3744.

Bours, R. J. E., Stuur, S., & Lankheet, M. J. M. (2007). Tuning for temporal interval in human apparent motion detection. *Journal of Vision*, *7*(1), 2. https://doi.org/10.1167/7.1.2

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research*, *14*(7), 519–527.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brownlow, S., Dixon, A. R., Egbert, C. A., & Radcliffe, R. D. (1997). Perception of movement and dancer characteristics from point-light displays of dance. *The Psychological Record*, *47*(3), 411-422.

Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H. J., & Rizzolatti, G. (2004). Neural circuits underlying imitation learning of hand actions: an event-related fMRI study. *Neuron*, *42*(2), 323-334.

Buehner, M. J. (2012). Understanding the Past, Predicting the Future: Causation, Not Intentional Action, Is the Root of Temporal Binding. *Psychological Science*, *23*(12), 1490–1497.

Buehner, M. J., & Humphreys, G. R. (2009). Causal Binding of Actions to Their Effects. *Psychological Science*, *20*(10), 1221–1228.

Buehner, M. J., & Humphreys, G. R. (2010). Causal Contraction: Spatial Binding in the Perception of Collision Events. *Psychological Science*, *21*(1), 44–48.

Bullock, M., & Gelman, R. (1979). Preschool children's assumptions about cause and effect: Temporal ordering. *Child Development*, 89-96.

Burr, D. C., Ross, J., & Morrone, M. C. (1986). Smooth and sampled motion. *Vision Research*, *26*(4), 643–652.

Caggiano, V., Fleischer, F., Pomper, J. K., Giese, M. A., & Thier, P. (2016). Mirror neurons in monkey premotor area F5 show tuning for critical features of visual causality perception. *Current Biology*, *26*(22), 3077-3082.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Carpenter, M., Call, J., & Tomasello, M. (2005). Twelve ‐ and 18 ‐ month ‐ olds copy actions in terms of goals. *Developmental Science*, *8*(1), F13-F20.

Cashon, C. H., Ha, O. R., Allen, C. L., & Barna, A. C. (2013). A U-shaped relation between sitting ability and upright face processing in infants. *Child Development, 84*(3), 802-809.

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences*, *104*(45), 17825-17830.

Chang, D. H., & Troje, N. F. (2008). Perception of animacy and direction from local biological motion signals. *Journal of Vision*, *8*(5), 3-3.

Chen, Y.-C., & Scholl, B. J. (2016). The Perception of History: Seeing Causal History in Static Shapes Induces Illusory Motion Perception. *Psychological Science*, *27*(6), 923–930.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Churchland, M. M., Priebe, N. J., & Lisberger, S. G. (2005). Comparison of the Spatial Limits on Direction Selectivity in Visual Areas MT and V1. *Journal of Neurophysiology*, *93*(3), 1235–1245.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*, 27755.

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*(8), 3563-3579.

Clarke, T. J., Bradshaw, M. F., Field, D. T., Hampson, S. E., & Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, *34*(10), 1171-1180.

Cohen, L. B., & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant behavior and development*, *21*(4), 713-731.

Cravo, A. M., Claessens, P. M. E., & Baldo, M. V. C. (2009). Voluntary action and causality in temporal binding. *Experimental Brain Research*, *199*(1), 95–99.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, *9*(5), 353-356.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55-67.

Darwin C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

Decety, J., & Grèzes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in cognitive sciences*, *3*(5), 172-178.

Desantis, A., & Haggard, P. (2016). Action-outcome learning and prediction shape the window of simultaneity of audiovisual outcomes. *Cognition, 153,* 33–42.

Desantis, A., Waszak, F., Moutsopoulou, K., & Haggard, P. (2016). How action structures time:

about the perceived temporal order of action and predicted outcomes. *Cognition*, *146*, 100-109.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*(8), 2051-2062.

Desrochers, S. (1999). Infants' processing of causal and noncausal events at 3.5 months of age. *The Journal of genetic psychology*, *160*(3), 294-302.

Dittrich, W. H., Troscianko, T., Lea, S. E., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception*, 25(6), 727-738.

Dodge, S., & Karam, L. (2017). Can the early human visual system compete with Deep Neural Networks?. *arXiv preprint arXiv:1710.04744*.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review*, *115*(1), 1.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470-2473.

Engbert, K., & Wohlschläger, A. (2007). Intentions and expectations in temporal binding. *Consciousness and Cognition*, *16*(2), 255–264.

Engbert, K., Wohlschläger, A., Thomas, R., & Haggard, P. (2007). Agency, subjective time, and other minds. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1261–1268.

Engemann, D. A., & Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, *108*, 328-342.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, *73*(6), 2608-2611.

Faro, D., Leclerc, F., & Hastie, R. (2005). Perceived causality as a cue to temporal distance. *Psychological Science*, *16*(9), 673-677.

Farrer, C., Valentin, G., & Hupé, J. M. (2013). The time windows of the sense of agency. *Consciousness and cognition*, *22*(4), 1431-1441.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1933-1941).

Fischl, B., Salat, D. H., Van Der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *Neuroimage*, *23*, S69-S84.

Fragkiadaki, K., Levine, S., Felsen, P., & Malik, J. (2015). Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4346-4354).

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. Cognitive brain research, *24*(1), 41-47.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593-609.

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *Journal of Neuroscience*, *32*(41), 14276-14280.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51*(7), 771-781.

Giese, M. A., & Poggio, T. (2003). Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, *7,* 267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86,* 446-460.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, *31*(5), 855-868.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). IEEE.

Grezes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. *Human Brain Mapping*, *12*(1), 1-19.

Gross, C. G., Bender, D. B., & Rocha-Miranda, C. E. D. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, *166*(3910), 1303-1306.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, *35*(6), 1167-1175.

Gunns, R. E., Johnston, L., & Hudson, S. M. (2002). Victim selection and kinematics: A point-light investigation of vulnerability to attack. *Journal of Nonverbal Behavior*, *26*(3), 129-158.

Haggard, P., Aschersleben, G., Gehrke, J., & Prinz, W. (2002). Action, binding and awareness. *Common Mechanisms in Perception and Action*, 266–285.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–385.

He, M., Walle, E. A., & Campos, J. J. (2015). A cross-national investigation of the relationship between infant walking and language development. *Infancy*, *20*(3), 283-305.

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature reviews neuroscience*, *9*(6), 467.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259.

Hirai, M., Fukushima, H., & Hiraki, K. (2003). An event-related potentials study of biological motion perception in humans. *Neuroscience Letters*, *344*(1), 41–44.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, *17*(1-3), 185-203.

Hume, D. (1888). *Hume's treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford, England: Clarendon Press. (Original work published 1739)

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, *104*(3), 427.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, *110*(2), 220.

Humphreys, G. R., & Buehner, M. J. (2009). Magnitude estimation reveals temporal binding at super-second intervals. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1542–1549.

Humphreys, G. R., & Buehner, M. J. (2010). Temporal binding of action and effect in interval reproduction. *Experimental Brain Research*, *203*(2), 465–470.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *science*, *286*(5449), 2526-2528.

Ionescu, C., Li, F., & Sminchisescu, C. (2011, November). Latent structured models for human pose estimation. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 2220-2227). IEEE.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence, 36*(7), 1325-1339.

Isik, L., Tacchetti, A., & Poggio, T. (2018). A fast, invariant representation for human action in the visual system. *Journal of Neurophysiology*, *119*(2), 631–640.

Jaeger, R. A. (1973). Action and subtraction. *The Philosophical Review, 82*(3), 320-329.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, *14*(2), 201-211.

Jokisch, D., Daum, I., Suchan, B., & Troje, N. F. (2005). Structural encoding and recognition of biological motion: evidence from event-related potentials and source analysis. *Behavioural brain research*, *157*(2), 195-204.

Karaminis, T., Turi, M., Neil, L., Badcock, N. A., Burr, D., & Pellicano, E. (2015). Atypicalities in perceptual adaptation in autism do not extend to perceptual causality. *PloS one*, *10*(3), e0120439.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kourtzi, Z. (2004). ‘But still, it moves.’ *Trends in Cognitive Sciences*, *8*(2), 47–49.

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & psychophysics*, *21*(6), 575-580.

Kozlowski, L. T., & Cutting, J. E. (1978). Recognizing the gender of walkers from point-lights mounted on ankles: Some second thoughts. *Attention, Perception, & Psychophysics*, *23*(5), 459-459.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011, November). HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision* (pp. 2556-2563). IEEE.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011, November). HMDB: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 2556-2563). IEEE.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299-321.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience*, *26*(11), 2894-2906.

Lappin, J. S., & Bell, H. H. (1976). The detection of coherence in moving random-dot patterns. *Vision Research*, *16*(2), 161–168.

Le, Q. V., Zou, W. Y., Yeung, S. Y., & Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR) 2011 IEEE Conference*. 3361–3368.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Leslie, A. M. (1982). The perception of causality in infants. *Perception, 11*(2), 173-186.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265-288.

Libertus, K., & Landa, R. J. (2013). The Early Motor Questionnaire (EMQ): A parental report measure of early motor development. *Infant Behavior and Development, 36*(4), 833-842.

Lin, F. H., Belliveau, J. W., Dale, A. M., & Hämäläinen, M. S. (2006). Distributed current estimates using cortical orientation constraints. *Human brain mapping*, *27*(1), 1-13.

Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016, October). Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision* (pp. 816-833). Springer, Cham.

Lu, H. (2010). Structural processing in biological motion perception. *Journal of Vision*, *10*(12), 13.

Lu, H., & Yuille, A. (2006). Ideal observers for detecting motion: Correspondence noise. *Advances in Neural Information Processing Systems*, *18*, 827-834.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review, 119*, 617–648.

Lu, H., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2006). Role of gamma-band synchronization in priming of form discrimination for multiobject displays. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 610.

Ludwig, W. (1953). Philosophical investigations. *London, Basil Blackwell*.

Manera, V., Del Giudice, M., Bara, B. G., Verfaillie, K., & Becchio, C. (2011). The Second-Agent Effect: Communicative Gestures Increase the Likelihood of Perceiving a Second Agent. *PLoS ONE*, *6*(7), e22650.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177-190.

Martinez, J., Black, M. J., & Romero, J. (2017). On Human Motion Prediction Using Recurrent Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4674–4683.

Masselink, J., & Lappe, M. (2015). Translation and articulation in biological motion perception. *Journal of Vision*, *15*(11), 10.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *258*(1353), 273-279.

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental psychology*, *31*(5), 838.

Meyers, E. (2013). The neural decoding toolbox. *Frontiers in neuroinformatics*, *7,* 8.

Michotte, A. E. (1963). *The perception of causality* (T. R. Miles, Trans.). London, England: Methuen & Co. (Original work published 1946)

Mikami, A., Newsome, W. T., & Wurtz, R. H. (1986). Motion selectivity in macaque visual cortex. II. Spatiotemporal range of directional interactions in MT and V1. *Journal of Neurophysiology*, *55*(6), 1328–1339.

Montepare, J. M., & Zebrowitz-McArthur, L. (1988). Impressions of people created by age-related qualities of their gaits. *Journal of personality and Social Psychology*, *55*(4), 547.

Moore, J. W., Lagnado, D., Deal, D. C., & Haggard, P. (2009). Feelings of control: contingency determines experience of action. *Cognition*, *110*(2), 279-283.

Moore, J., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and cognition*, *17*(1), 136-144.

Morgan, M. J., & Ward, R. (1980). Conditions for motion flow in dynamic visual noise. *Vision Research*, *20*(5), 431–435.

Nadel, S., & Poss, J. E. (2007). Early detection of autism spectrum disorders: screening between 12 and 24 months of age. *Journal of the American Academy of Nurse Practitioners*, *19*(8), 408-417.

Neri, P., Luu, J. Y., & Levi, D. M. (2006). Meaningful interactions can enhance visual discrimination of human agents. *Nature Neuroscience*, *9*(9), 1186–1192.

Newsome, W. T., Mikami, A., & Wurtz, R. H. (1986). Motion selectivity in macaque visual cortex. III. Psychophysics and physiology of apparent motion. *Journal of Neurophysiology*, *55*(6), 1340–1351.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In I*EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436.

Nishitani, N., & Hari, R. (2000). Temporal dynamics of cortical representation for action. *Proceedings of the National Academy of Sciences*, *97*(2), 913-918.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5*(2), 193-207.

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex*, *22*(5), 981-995.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & Psychophysics*, *62*(5), 889–899.

Pavlova, M., Lutzenberger, W., Sokolov, A., & Birbaumer, N. (2004). Dissociable cortical processing of recognizable and non-recognizable biological movement: analysing gamma MEG activity. *Cerebral Cortex*, *14*(2), 181-188.

Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, *93*(1), 603-608.

Peelen, M. V., & Downing, P. E. (2007). The neural basis of visual body perception. *Nature reviews neuroscience*, *8*(8), 636.

Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, *49*(6), 815-822.

Peng, Y., Javangula, P. R., Lu, H., & Holyoak, K. J. (2018). Behavioral oscillations in verification of relational role bindings. In C. Kalish, M. Rau, J. Zhu & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal Action: A Fundamental Constraint on Perception and Inference About Body Movements. *Psychological Science, 28*(6), 798–807.

Phillips, S., Takeda, Y., & Singh, A. (2012). Visual feature integration indicated by phase-locked frontal-parietal EEG signals. *PLoS One*, *7*(3), e32502.

Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: a functional MRI study. *Proceedings of the National Academy of Sciences*, *102*(19), 6996-7001.

Price, H. (1992). Agency and causal asymmetry. *Mind, 101*, 501–520.

Rakison, D. H., & Krogh, L. (2012). Does causal action facilitate causal perception in infants younger than 6 months of age?. *Developmental science, 15*(1), 43-53.

Rankin, M. L., & McCormack, T. (2013). The temporal priority principle: at what age does this develop? *Frontiers in Psychology, 4*(178).

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., ... & Zemel, R. S. (2018). Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, *21*(5), 188-194.

Rizzolatti, G., & Fadiga, L. (1998). Grasping objects and grasping action meanings: The dual role of monkey rostroventral premotor cortex (area F5). *Sensory Guidance of Movement*, *218*, 81-103.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661.

Rohde, M., Greiner, L., & Ernst, M. O. (2014). Asymmetries in visuomotor recalibration of time perception: Does causal binding distort the window of integration? *Acta Psychologica, 147*, 127–135.

Rohde, M., Scheller, M., & Ernst, M. O. (2014). Effects can precede their cause in the sense of agency. *Neuropsychologia*, *65*, 191-196.

Runeson, S., & Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender recognition, and deceptive intention. *Journal of experimental psychology: general*, *112*(4), 585-615.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review, 120*, 411–437.

Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science*, *13*(6), 493-498.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*, 299–309.

Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience*, *25*(47), 11055-11059.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139-159.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and brain sciences*, *16*(3), 417-451.

Shiffrar, M., & Freyd, J. J. (1990). Apparent Motion of the Human Body. *Psychological Science, 1*(4), 257–264.

Shiffrar, M., & Freyd, J. J. (1993). Timing and Apparent Motion Path Choice With Human Body Photographs. *Psychological Science, 4*(6)*,* 379–384.

Shirai, N., & Imura, T. (2014). Implied motion perception from a still image in infancy. *Experimental Brain Research*, *232*(10), 3079–3087.

Shirai, N., & Imura, T. (2016). Emergence of the ability to perceive dynamic events from still pictures in human infants. *Scientific Reports*, *6*(1).

Shu, T., Thurman, S., Chen, D., Zhu, S.-C., & Lu, H. (2016). Critical features of joint actions that signal human interaction. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 574–579).

Sigman, E., & Rock, I. (1974). Stroboscopic Movement Based on Perceptual Intelligence. *Perception*, *3*(1), 9–28.

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences (USA)*, *105*(2), 809-813.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568-576).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *In Proceedings of international conference on learning representations (iclr)*.

Singh-Curry, V., & Husain, M. (2009). The functional role of the inferior parietal lobe in the dorsal and ventral stream dichotomy. *Neuropsychologia*, *47*(6), 1434-1448.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407-419.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology*, *46*(1), 129.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*(1), 1–10.

Spiridon, M., Fischl, B., & Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human brain mapping*, *27*(1), 77-89.

Spröte, P., Schmidt, F., & Fleming, R. W. (2016). Visual perception of shape altered by inferred causal history. *Scientific Reports*, *6*(1).

Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, *51*(5), 651-659.

Straube, B., & Chatterjee, A. (2010). Space and time in perceptual causality. *Frontiers in Human Neuroscience*, *4*, 28.

Strickland, B., & Keil, F. (2011). Event Completion: Event Based Inferences Distort Memory in a Matter of Seconds. *Cognition, 121*(3), 409–415.

Su, J., & Lu, H. (2017). Flash-lag effects in biological motion interact with body orientation and action familiarity. *Vision Research*, *140*, 13–24.

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception*, *13*(3), 283-286.

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017-1024).

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017-1024).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks*. arXiv preprint arXiv:1312.6199.

Theusner, S., de Lussanet, M., & Lappe, M. (2014). Action recognition by motion detection in posture space. *Journal of Neuroscience, 34*(3), 909-921.

Thurman, S. M., & Lu, H. (2013). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychological Science*, *24*, 1133–1141.

Thurman, S. M., & Lu, H. (2014a). Perception of social interactions for spatially scrambled biological motion. *PLoS ONE, 9*(11), Article e112539. doi:10.1371/journal.pone.0112539

Thurman, S. M., & Lu, H. (2014b). Bayesian integration of position and orientation cues in perception of biological and non-biological forms. *Frontiers in human neuroscience, 8*.

Thurman, S. M., & Lu, H. (2016a). Revisiting the importance of common body motion in human action perception. *Attention, Perception, & Psychophysics*, *78*(1), 30-36.

Thurman, S. M., & Lu, H. (2016b). A comparison of form processing involved in the perception of biological and nonbiological movements. *Journal of vision, 16*(1), 1-1.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current Biology*, *16*(8), 821-824.

Tsang, T., Ogren, M., Peng, Y., Nguyen, B., Johnson, K. L., & Johnson, S. P. (2018). Infant perception of sex differences in biological motion displays. *Journal of Experimental Child Psychology, 173,* 338-350.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature neuroscience*, *6*(9), 989.

Tse, P., Cavanagh, P., & Nakayama, K. (1998). The role of parsing in high-level motion processing. *High-Level Motion Processing: Computational, Neurobiological, and Psychophysical Perspectives,* 249–266.

Tucciarelli, R., Turella, L., Oosterhof, N. N., Weisz, N., & Lingnau, A. (2015). MEG multivariate analysis reveals early abstract action representations in the lateral occipitotemporal cortex. *Journal of Neuroscience*, *35*(49), 16034-16045.

Urgesi, C., Candidi, M., Ionta, S., & Aglioti, S. M. (2007). Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. *Nature neuroscience*, *10*(1), 30.

van Boxtel, J. J., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. PLoS One, 7(5), e37085.

van Boxtel, J. J., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of vision*, *13*(12), 7-7.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*(3), 564-584.

VanRullen, R., & Dubois, J. (2011). The psychophysics of brain rhythms. *Frontiers in psychology*, *2*, 203.

Virji-Babul, N., Cheung, T., Weeks, D., Kerns, K., & Shiffrar, M. (2007). Neural activity involved in the perception of human and meaningful object motion. *Neuroreport*, *18*(11), 1125-1128.

von der Malsburg, C., & Buhmann, J. (1992). Sensory segmentation with coupled neural oscillators. *Biological cybernetics*, *67*(3), 233-242.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222–236.

Walle, E. A., & Campos, J. J. (2014). Infant language development is related to the acquisition of walking. *Developmental Psychology*, *50*(2), 336.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision* (pp. 20-36). Springer International Publishing.

Wegner, D. M. (2004). Précis of the illusion of conscious will. *Behavioral and Brain Sciences*, *27*(5), 649-659.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American psychologist*, *54*(7), 480.

Wertheimer, M. (1912). Experimentelle Studien uber das Sehen von Bewegung. *Zeitschrift Fur Psychologie, 61*.

White, N. C., Fawcett, J. M., & Newman, A. J. (2014). Electrophysiological markers of biological motion and human form recognition. *Neuroimage*, *84*, 854-867.

White, P. A. (1999). Toward a causal realist account of causal understanding. *American Journal of Psychology*, *112*, 605-642.

White, P. A. (2005a). Visual causal impressions in the perception of several moving objects. *Visual Cognition*, *12*(2), 395-404.

White, P. A. (2005b). Visual impressions of interactions between objects when the causal object does not move. *Perception*, *34*(4), 491-500.

White, P. A. (2006). The causal asymmetry. *Psychological Review, 113*, 132–147.

White, P. A., & Milne, A. (1997). Phenomenal causality: Impressions of pulling in the visual perception of objects in motion. *The American journal of psychology*, *110*(4), 573-602.

White, P. A., & Milne, A. (1999). Impressions of enforced disintegration and bursting in the visual perception of collision events. *Journal of Experimental Psychology: General*, *128*(4), 499-516.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The Perceived Onset Time of Self- and Other-Generated Actions. *Psychological Science*, *14*(6), 586–591.

Woods, A. J., Hamilton, R. H., Kranjec, A., Minhaus, P., Bikson, M., Yu, J., & Chatterjee, A. (2014). Space, Time, and Causality in the Human Brain. *NeuroImage*, *92,* 285–297.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*(1), 73-77.

Woodward, A.L., & Guajardo, J.J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development*, *17*, 1061–1084.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624.

Young, M. E. (1995). On the origin of personal causal theories. *Psychonomic Bulletin & Review*, *2*(1), 83-104.

Young, M. E., Rogers, E. T., & Beckmann, J. S. (2005). Causal impressions: Predictingwhen, not justwhether. *Memory & Cognition*, *33*(2), 320-331.

Zalla, T., Labruyère, N., Clément, A., & Georgieff, N. (2010). Predicting ensuing actions in children and adolescents with autism spectrum disorders. *Experimental Brain Research*, *201*(4), 809-819.