The History of Neuroscience in Autobiography
Volume 13

Edited by Thomas D. Albright and Larry R. Squire
Published by Society for Neuroscience
ISBN: 978-0-916110-12-3

Shimon Ullman

pp. 436–471

https://doi.org/10.1523/hon.013011

Shimon Ullman

Born:

Jerusalem, Israel
January 28, 1948

Education:

Hebrew University Jerusalem, BSc, Mathematics, Physics, Biology, summa cum laude (1973)
Massachusetts Institute of Technology, PhD, Electric Engineering and Computer Science (1977)

Appointments:

Research Associate, AI Lab, MIT (1977–1979)
Principal Research Scientist, AI Lab, MIT (1979–1980)
Associate Professor, Brain & Cognitive Sciences, MIT (1981–1985)
Associate Professor of Computer Science, Weizmann Institute (1981–1985)
Professor, Brain & Cognitive Science, MIT (1986–1993)
Professor of Computer Science, Weizmann Institute (1985–present)

Honors and Awards (Selected):

The David. E. Rumelhart Prize in Human Cognition (2008)
Cognitive Science Society Fellow (2008)
Member, Israeli Academy of Sciences and Humanities (2011)
The Israel Emet Prize for Art, Science and Culture (2014)
The Israel Prize in Computer Science (2015)
Foreign member, American Academy of Arts and Sciences (2016)
A. Rosenfeld Award for lifetime achievement in computer vision (2019)

Shimon Ullman used computational modeling, at both functional and network levels, to study information processing in the cortical visual stream. His studies combine computational, psychophysics, and collaborations in biological experimentations. His earlier work focused on the topics of motion perception and object recognition. In the motion area, he established conditions under which a three-dimensional structure can be recovered from dynamic scenes, and he developed the first method for recovery of structure from motion. In object recognition, his early models were the first to combine bottom-up with top-down segmentation, and to localize objects and their parts and subparts in a single bottom-up, top-down cycle. Moving from objects to complete scenes, Ullman developed a model that performs human-like scene interpretation, using iterative bottom-up and top-down processing in a "counter-streams" structure motivated by cortical circuitry. Using data from visual learning by infants, Ullman developed models that can learn complex concepts from dynamic visual scenes, without supervision or annotated data. This modeling shows how innate domain-specific "proto concepts" can guide the visual system to acquire meaningful concepts and reveals how a rich conceptual system can gradually arise from the combination of innate mechanisms and visual experience.

Shimon Ullman

Stormy Beginning

I was born in the midst of a raging war, in a makeshift facility in downtown Jerusalem. Only three hours later, I was carried with my mother in an armored vehicle to the well-equipped Hadassah Mount Scopus Hospital, just outside the city, to spend the night and the following day. The hospital had to close down its operation a couple of months later, following the Hadassah medical convoy massacre, when a convoy carrying supplies to the hospital was ambushed. Only a few of the medical staff in the convoy survived the attack, among them my father, Theodor David Ullman ("Theo" for us), who was a physician in Hadassah. He found shelter in a ditch not far from the road and hid there until night time. He then made his way to our apartment in Jerusalem, avoiding the Arab villages along the way, and arriving home by dawn. My mother heard about the fate of the Hadassah convoy from the radio news, and I can only imagine, since she never described to me, how she felt when Theo appeared at the doorstep in the early morning hours.

The apartment we lived in was in a place called the "Mandelbaum gate," which ended up in a no-man's-land between Israel and Jordan. During the 1948 war, it was very close to the frontline and became a place of increasing hostilities. It was hit by snipers' bullets several times, and at one point, a larger projectile landed inside our apartment. When this happened, my parents carried the children, my sister Edna and me, and left the apartment, leaving all the belongings behind, never to come back.

I recalled these events many years later, when I was standing with my wife Chana in a demonstration in the Palestinian neighborhood of Sheikh Jarrah in East Jerusalem. The demonstration was in support for Palestinian residents who were being evacuated from apartments they had moved into during the 1948 war. The place where the demonstration took place was about midway between the apartment we had to leave during the war and the place where the Hadassah convoy was ambushed, a short walking distance from both. The past memories and the demonstration formed an eerie mix, underscoring for me in a personal way the need to resolve the ongoing Israeli-Palestinian conflict.

Family and Early Life

My father was a German Jew, born in Würzburg to a large orthodox family. He studied medicine in Würzburg and Berlin, and came to Israel in 1934. He was a physician in Hadassah hospital and a professor at the Hebrew University in Jerusalem, specializing in nephrology. After spending a year at

the Mount Sinai Hospital in New York, he was the first to introduce kidney dialysis to Israel. I remember him telling me about the ethics committee established in the hospital, which had to make painful decisions regarding who would be eligible for dialysis treatment, because it was impossible to treat all the patients needing treatment. He was a brilliant thinker, was quiet and hard-working, and had immense general knowledge.

My mother, Lisa Findler, was born and raised in Vienna, until the age of 16. She left after the Anschluss of 1938 and went on her own via Italy to Palestine. Her parents were supposed to join a few weeks later, but never made it out. After finishing British matriculation exams by correspondence, she enrolled into a nursing school and became a nurse in Hadassah Hospital, where she met my father. She kept working as a nurse and later as a laboratory assistant, until I left home years later, and then started academic studies at the Hebrew University, in the classics department. She eventually became a staff member in the university, teaching Greek and Latin, and became a scholar and a translator. At the age of 77, she undertook the daunting task of translating from the Greek Josephus's first-century book *The War of the Jews*. This is a major book, described by the historian Steve Mason as "perhaps the most influential non-biblical text of Western history," and one of the first books printed by Gutenberg in the 15th century. The translation took 10 years of hard work, and when it was published, it was an immediate literary and commercial success. I remember that in one of his visits to Israel, Eric Kandel was at our house and met my mother. He was delighted to meet a fellow Viennese who, like him, was interested and highly knowledgeable in literature, art, and music.

My sister, Edna, was only a year and a half older than me. We have been very close, and her early death in 2010 was an immense, life-transforming loss for me. She was a philosopher and was interested in issues of rationality as well as decision-making problems, from simple to transformational, and problems of social order. The issues she was interested in and worked on had bearing not only on philosophy but also on political science, psychology, sociology, cognitive science, economics, law, and public policy. Her personal life echoed her philosophy; she was a highly rational person, and at the same time, she was passionate about the values she studied and believed in. She was the chair of the Association for Civil Rights in Israel and on the management of the New Israel Fund and of B'Tselem (The Israeli Information Center for Human Rights in the occupied territories). She was also a gifted translator, in both science and literature. She translated to Hebrew Wittgenstein's *Philosophical Investigations* and *On Certainty* from the original German, and her wonderful poetry translations from both German and English appeared in Haaretz's literary supplements. I think one can get a wonderful glimpse of who Edna was from a *New Yorker* piece called "The Unmothered," written by her daughter Ruth Margalit (https://www.newyorker.com/books/page-turner/the-unmothered).

During my school days, both elementary and high school, I was an avid reader, taking books out of the school library several times a week and enjoying reading them. I now have grandchildren who do the same, and I enjoy seeing their enthusiasm and watching how reading enriches their inner worlds. In high school, I became highly interested in science, reading well beyond the material covered in school. I was fortunate to have an inspiring teacher of mathematics, a young woman in her twenties who had the gift of conveying the beauty and rigor of mathematical thinking.

When high school ended, it was time to enlist to military service, which is compulsory in Israel. It was clear to me by that time that my future would be in academic research, but I decided to spend my service doing something completely different and joined the air force for more than five years. This was an eventful and challenging period, but I will not elaborate on it, as this period was about as far from an academic path as one can imagine.

Starting Academic Life: The Hebrew University

After a long break from any form of academic studies, I was eager to start my scientific education. At the Hebrew University, you could either focus on a single field, such as physics, mathematics, or biology, or combine two fields, taking half of your academic credits in each field. I knew I wanted to take mathematics as a major field, but after some deliberations, I decided to combine mathematics and physics. Math was my major field, and I joined an accelerated track that had an expanded curriculum. The beginning was demanding, as the academic year started before I was released from service, and it took a while before I could devote my full time to the academic studies. This did not detract from my excitement, and I thoroughly enjoyed delving deeply into math and physics. Math was particularly rewarding, as I was taught by brilliant mathematicians and teachers, such as Hillel Furstenberg, Azriel Levy, and Yakar Kannai, who were able to get across not only mathematical knowledge but also their personal ways of thinking about mathematics. Regarding physics, although I enjoyed it, I was not sure whether this was in fact the field of science I wanted to get into. I therefore attended during the first year a couple of classes in neuroscience, to learn a bit about the brain, which I thought was a fascinating topic and on which I knew relatively little. At the end of the first year, I found myself in a Frostian dilemma in "the road not taken," facing diverging roads, and being "sorry I could not travel both, and be one traveler." It felt like a major crossroads, choosing between exploring the external universe of physics, and the internal universe of the brain, thought processes, and the subjective world of awareness and consciousness. Like many big decisions, this was not a utilitarian choice but rather an opportunity to imagine what would be for me more exciting and rewarding. It took some time, but I chose the internal universe, and although I find the big questions of physics fascinating, I did

not regret the choice. The practical outcome of these deliberations was that I added a full program in neurobiology, without giving up physics, to my undergraduate program.

Toward the end of my undergraduate studies, I started to have a general notion of the kind of basic scientific problems I would like to pursue. I wanted to study how the brain works, by somehow combining aspects of biology with some form of mathematical or theoretical work. I was more attracted toward the theoretical side, but I wanted to stay close to the actual anatomy and physiology of the brain. On the theoretical side, I was starting to think about models that would be able to mimic some of the functions of the human brain. I was wondering if somewhere there were research groups or individuals working along such lines. Searching for relevant information was not as straightforward as it is today, but at some point, I came across descriptions of the artificial intelligence (AI) laboratory at Massachusetts Institute of Technology (MIT), and I was highly intrigued. There were reports, many in the form of internal AI memos, describing attempts to develop computer models that were trying to replicate some forms of human thinking. At the same time, MIT had a strong group of neuroscientists, housed at the time at MIT's psychology department, headed by Hans-Lukas Teuber. This appeared to me to be a wonderful unique opportunity, to study at MIT, combining neuroscience with the research laboratory with the strange name of AI. I applied, describing my interests, and when I was accepted, the decision was immediate: I would go to MIT to follow these research interests.

During my undergraduate studies at the Hebrew University, I met my future wife, Chana. We were introduced by a classmate of my sister, who told Chana "you must meet Edna's little brother." We met, and for me, it was an immediate click. She was (still is) beautiful and highly intelligent, with a unique personality, sensitive, empathic, original, and perceptive. When it was time for me to leave for MIT to start the school year, she was in the last stages of finishing her master of science degree in psychology at the Hebrew University. We decided that I would go initially on my own, and she would join me a few months later, as soon as she hands in her thesis. In late August 1973, I flew to Boston to start my MIT studies. We did not know then that we would meet again sooner than expected, under strange circumstances.

A Student at MIT: Human and Computational Vision

I settled down in Cambridge, and started the school year, when life took a sudden turn, without a warning. It was an early Saturday morning of October 6, 1973, when breaking news started to broadcast stories about intense fighting raging between Israel and its neighbors, Egypt on the south and Syria on the north. There was initially considerable confusion, and it was unclear whether these were border skirmishes or an all-out war. I was

trying to get more information in a series of phone calls with friends in Israel and in the Boston area, and with the Israeli embassy. The situation began to look serious enough, and I decided to fly home immediately. Flights had been suspended, but there was one last El Al flight leaving New York's Kennedy airport in the afternoon, and the embassy informed me that there was a seat reserved from me on this flight. I took the car I just purchased in Cambridge, and drove to JFK as fast as I could to catch the flight. With me, I took a fellow student at MIT, Benjamin (Bibi) Netanyahu. I knew his brother, Yoni (later killed while leading the rescue raid at Entebbe Airport in Uganda), who told me to look up his brother who was already an architecture student at MIT.

When we arrived in Israel (the stewardess wished us "a pleasant stay in Israel"), I took a cab to my base, where I was greeted by my squadron commander, Udi. Good natured, and always calm, he told me "have a good night's sleep; you will have a busy day tomorrow." The next day, Monday, was indeed a busy day, during which Udi was unfortunately shot down. I was supposed to have my first exam at MIT that day, but instead, I found myself over the Suez Canal in the morning, and over the Golan Heights in the afternoon.

When the war was over I came back to MIT, and a few days later I met David Marr, a meeting that had a profound impact on my academic life. Within a short time, he became my academic mentor, and later also a close and dear friend. David arrived at MIT only a few months before me, from Trinity College in Cambridge, England. I was struck by his scientific thinking and intelligence, and by the close fit between what I was looking for, and what he was already doing in his research. Initially, he could not be my formal doctoral supervisor, because his position was of a researcher rather than faculty member. Marvin Minsky, who was then director of the MIT AI lab, graciously agreed to act as my formal supervisor, but let David do the actual supervision. David already had a brilliant scientific record from his Cambridge doctorate and subsequent work. During this work, he produced a mathematical theory of three major parts of the brain: the archicortex, cerebellum, and neocortex. This work, in particular the cerebellum theory, had a large impact on the neuroscience community. David became dissatisfied with the general kind of modeling developed in his own work, however, and started to form a new modeling direction he wanted to pursue.

This new direction took several years to develop and take form, but the main ingredients were clear from the beginning. A basic aspect of the new direction was the combination of brain and computation. This meant combining empirical studies of the brain with computational models that could carry out functions performed by the brain. In the domain of vision, the computational challenge was to create computer models that could carry out tasks that our visual system performs for us and to use the modeling to analyze and interpret empirical findings as well as to guide new experiments. Many of the functions carried out by our visual system are complex

and difficult to imitate by computational models. Constructing successful computational models is therefore likely to reveal useful information about the nature of the task and about possible processes that the visual system may be using to carry out the tasks in question. Marr realized from his own work how difficult it is to go from empirical data obtained by neuroscientists to an understanding of how the brain functions. Computational modeling offered a complementary way to gather relevant information, and combining the two sources of information became the foundation of the emerging field of computational neuroscience. This potential integration motivated Marr to move to MIT's AI lab, because this was a research laboratory where researchers where trying to develop functioning brain-like computational models. My own motivation in coming to MIT was similar, and I was fortunate to meet Marr upon my arrival to the AI lab. It was true that MIT AI researchers were trying to develop computational models with some capacities of the human brain, but not much attention has been paid to possible relationships between the models and the human brain. Luckily, when I arrived, the AI lab already had the most qualified person in the world to lead this research direction.

Because my knowledge of human vision was limited, I spent significant time reading and studying the anatomy, physiology, and psychophysics of the primate visual system. MIT had wonderful people to learn anatomy and physiology from, and I learned a lot from them, in particular Mike Sryker (see volume 11) and Peter Schiller (see volume 7). Anatomy looked to me initially somewhat dry and boring, but I gradually learned to appreciate its importance and even elegance. At some point during this early period, I went with Marr to visit Francis Crick at the Salk Institute, who became highly interested in the brain in general and the visual system in particular. He knew David from their Cambridge days and invited him often to learn from him and discuss novel ideas and approaches. I was struck by Crick's emphasis on the anatomy of the visual system. He became an expert on the detailed structure of the visual cortex, and he even constructed some three-dimensional models of parts of the system. This was perhaps related in part to his DNA work in which structure and function were of course intimately related, and he similarly used cortical structures as a source of insight about functional mechanisms.

Along with getting more familiar with the visual system, I started to think about different visual functions, and what it would take to perform them by computational models. I wanted the model to be able to approximate human performance, but also to consider possible implementations in biologically plausible network models and possibly comparisons with neurophysiological or psychophysical data. In trying to model visual tasks, I soon learned to appreciate the complexity and sophistication of the visual process. Even for seemingly simple visual tasks, obtaining a model that can rival human performance can be surprisingly challenging. I will describe

briefly an example of a visual problem I was concerned with early on in my studies, which can illustrate some aspects of the state of the art at the time, and early developments in dealing with computational vision.

The problem was to characterize the shape of subjective contours generated by the visual system and to provide a model for their generation. These contours do not exist objectively in the image, but they are generated by our visual system under some conditions that suggest the existence of an invisible surface occluding the background. There has been considerable work on the conditions that make the visual system infer, or suggest, the existence of such invisible surface. However, there was no empirical or theoretical analysis of the shape of the subjective contours generated by the visual system. Because subjective contours are generated by the visual system, the analysis of their shape may provide clues about the mechanisms that generate them. The theoretical analysis suggested that the shape of a contour that fills in the gap between two boundary edges is composed of the arcs of two circles, tangent to the boundary edges, meeting smoothly, and minimizing the total curvature along the contour. This shape was derived from a number of general global assumption made about the contours, based on empirical observations (isotropy, smoothness, locality, and minimal curvature). If these properties are assumed, the two-arc shape of the entire contour can be derived analytically. This conclusion about the circular arcs as elements of the full contour is also consistent with later work about the facilitation in the primary visual cortex between cocircular contour segments [1]. As my first attempt in network modeling, a simple network model was proposed for generating the full subjective contours given the visible edge fragments. The model is based on local interaction between neighboring line elements that depend on their orientation difference, and eventually selecting the most active line element at each point.

In related work some 10 year later, the model for subjective contours was extended to a model that constructs a saliency map, which is a representation of the image emphasizing salient locations [2,3]. This work was done together with my then-student Amnon Shashua, who later on, as a professor at the Hebrew University, became the cofounder of Mobileye, a leader of the autonomous cars industry. Amnon was a superb student, sharp and creative, and it was clear early on that he has an unusual combination of scientific talents together with a practical sense and leadership qualities. When he finished his master of science work at the Weizmann Institute, he moved to MIT to do his doctorate with Tommy Poggio (see volume 8).

The Interpretation of Visual Motion

The main part of my doctoral research was concerned with visual motion—that is, how the visual system measures the velocity of objects in the image, and how objects' motion is used to recover their three-dimensional shape.

I was fascinated by the ability of the human visual system to recover the three-dimensional shape of moving objects based on motion information alone, even when each static image contained no information about the objects' shape. Consider, for example, a cloud of points moving together rigidly in space. A single projection of this configuration onto an image produces just a random-looking arrangement of image points, which does not reveal the three-dimensional structure of the actual object. However, as soon as the object starts to move, the dynamic pattern of the moving object gives rise to a vivid perception of the true three-dimensional object. This has been a long-known phenomenon in the study of visual perception, called the "kinetic depth effect," first described by the great experimental psychologist Hans Wallach and coworkers in the 1950s. In the perception literature, it was sometimes explained in terms of a "tendency" of the perceptual system to perceive rigid objects. But behind this "tendency" lies a striking capacity of the visual system to solve a complex problem: given a collection of points moving on a plane, the visual system is able to determine whether they can come from the projection of an object (or perhaps more than a single one) moving rigidly in space and then recover the three-dimensional configuration of the points in space.

I referred to this recovery as the "structure from motion" computation and was interested in both the computational and empirical aspects of this visual capacity. The computational challenge was to understand the conditions under which the problem can be solved and to develop a model that could perform the task: given a dynamic image, it will be able to identify whether it could originate from the projection of a three-dimensional object, or perhaps several objects, moving in space, and recover their three-dimensional shape. It was unclear at the time whether the recovered structure was unique or whether there were always multiple possible three-dimensional shapes and motions, which were all compatible with the observed changing image. On the empirical side, the goal was to study psychophysically the properties of humans' perception of structure from motion. Another long-term goal was to eventually identify and understand the brain mechanisms involved in this visual capacity.

I examined the computational problem in detail, under somewhat different conditions (such as objects close to the observer, or more distant objects, where perspective effects are weak). For the case of distant objects, it turned out that three views of four non-coplanar points are sufficient to guarantee a unique three-dimensional structure. The uniqueness is in fact up to a depth-reversal (reflection about the frontal plane). This ambiguity is inherent, because the parallel projection of a rotating object is identical to the projection of the reversed object rotating in the opposite direction. My initial analysis and method for recovering the object used five points, and an extension by D. Fremlin (in a personal communication) reduced the configuration to four points. Following this analysis, uniqueness results and

shape-extraction methods have been extended to a variety of conditions. An important development was the formulation by Longuet-Higgins of an algorithm for recovering the structure of objects based on two perspective projections using the notion of a "fundamental matrix," which became a standard in the field [4]. The recovery of structure from motion developed to become a rich field, going in a variety of directions, and producing over the years advanced and sophisticated methods for shape recovery.

What is the relation between these computational results and human perception of structure from motion? It turns out that the shape recovery method based on the computational analysis was qualitatively consistent on a range of properties with perceptual phenomena discovered in empirical studies. There was, however, one difference that I found intriguing, where human perception performed better than the computational scheme. Human perception can also recover the structure of moving objects that are not entirely rigid but that distort and change while they are moving. Dealing with such distorting objects suggested a novel scheme for the recovery of structure from motion, which had another advantage: it appeared to be more appealing from the point of view of biological plausibility. The modified scheme did not recover the object's shape by solving some set of mathematical equations. Instead, it proposed an initial approximation to the shape and continuously improved the solution as more data became available when the object continued to move [5].

The recovery of structure from motion is an impressive achievement of the visual system and its use of motion information, and it was the first problem in the area of visual motion I was attracted to. But the first problem faced by a vision system that analyzes visual motion is the surprisingly complex problem of measuring visual motion. The motion of elements and regions in the image is not given directly, but must be computed from more elementary measurements. The initial registration of light by the eyes is in terms of light intensity and its changes over time. From this, the visual system extracts measurements of the direction and velocity of motion across the visual field. Extracting such motion information with sufficient detail and accuracy that will allow, for example, the recovery of structure from motion, turned out to be a challenging task, and even today the problem is not yet entirely solved. Based on past work, I suggested that the visual system uses two complementary methods for extracting visual motion, one is intensity based, and the second uses localized feature matching to estimate the motion [6].

In the doctoral work, I focused on the feature-matching method, because it is more directly relevant to the extraction of structure from motion. The study of visual motion, including the measurement of visual motion and the recovery of structure from motion became the subject matter of my doctoral work, completed in 1977, and published in book titled *The Interpretation of Visual Motion* [7].

During my doctoral period, Marr and I gradually became close friends. He was amazing, both in terms of scientific guidance and as a human being. Working with David was always challenging, exciting, and rewarding. It was hard work, but it was a lot of fun. I had the pleasure of interacting with, and learning from, a number of other faculty members, in particular Whitman Richards in the psychology department and Patrick Winston in the AI lab. Whitman was a rich source of knowledge about human vision and was always willing to entertain new and unconventional views. Patrick was my main source of learning about AI, combining broad knowledge with insightful intuitions. Both Whitman and Patrick realized and appreciated from the very start David's vision and the value of the program he started at MIT.

Another colleague who was central from the beginning in the work with Marr, and quickly became a close personal friend, was Tommy Poggio. Tommy visited the AI lab in 1973 and quickly became a scientific collaborator and close personal friend of David. Marr was highly impressed by Tommy's work and by his outstanding scientific qualities. I remember David telling me about the exciting work of Poggio on the fly's visual system, which he conducted at the Max-Planck Institute in Tübingen. Eventually, Marr convinced Tommy to join MIT on a permanent basis, and Tommy and his wife Barbara and their two children relocated to Boston. I remember this period very warmly; it was a wonderful time both scientifically and in terms of personal friendships. The three of us became close friends, spending time together in scientific discussions, outside work, at dinners, and on vacations with our families. At some later time, as the work progressed, we entertained the idea of starting together some sort of an entity, a consulting group or a company, exploring potential applications of computer vision. We gave it the name Cambridge Intelligent System and had stationary with this title printed over it, but we never got much beyond this point. Unfortunately, this period did not last very long, and our lives came to a sharp turn when David became seriously ill. David's illness came as a shock. He did not feel well one evening in December 1977 and was admitted to MIT's infirmary to undergo some tests. He called me from the MIT infirmary the next day and asked me to close my office door. He then said briefly and without any introduction that he was just diagnosed with acute leukemia.

The period that followed was very painful. What made it bearable, and for a meaningful time even a happy period, was David's meeting with Lucia Vaina and their falling in love. They met a few months after David's initial diagnosis and became very close within a short time, despite the ominous uncertainty. Lucia's love and support helped David through the most difficult times. For extended periods, he maintained an almost normal routine, doing intensive scientific work, some of it in collaboration with Lucia. Twice during his illness, we thought that there was some hope. The first was during his first remission. Everyone hoped that perhaps, by some miracle, the disease would not come back. We took a vacation together in Vermont, and he

resumed his work with his usual intensity. After a period, he felt weaker and went to the hospital for tests. He came to my office to call the hospital about the tests results and found out that it was indeed a relapse. We sat in my office for a long time devastated by the news. The second hope came when a physician in Cambridge, England, had some initial success with a vaccine against leukemia. David was hospitalized in Addenbrooke's hospital in Cambridge. He was very weak and worked on his book. When I came to visit, I met the physician, who was very supportive and promised to help as much as he could. When David came back to the United States, Tommy Poggio managed to bring some of the Cambridge vaccine with him, but the vaccine did not work for David (and it did not prove effective in later clinical trials).

The final period, when David already suspected that the battle was lost, was in fact a quietly happy one. He was happily married to Lucia and was working intensively on his book and a number of other projects. In his premature death, the scientific world lost an intellectual giant who, in the short time, made a huge impact on his field. I lost a warm, brilliant, exciting, and unusual friend.

During the doctoral time, on the last day of 1974, Chana and I got married. The wedding took place in the MIT chapel. This is a charming and nondenominational place on campus, which can be quickly transformed to fit any type of service. The wedding was a small and very pleasant event, with close family members, old friends, and some new friends we met in Cambridge, and who remained friends to this day. Shortly after our marriage, Chana enrolled in the doctoral program in psychology at Boston University, and after graduation she started her career as a clinical psychologist.

Researcher at the AI Lab

Following my doctorate, I had a research position at the AI lab for three years, first as a research associate and then as a principal research scientist. My MIT mentors, David and Whitman Richards, thought that this would be a good way to continue research at MIT before applying for an MIT faculty position.

During my doctoral research, I found a range of open problems related to visual motion I wanted to look into, and I continued to study aspects of visual motion well beyond the doctoral period. A study of visual motion from 1981 was one of the only two papers Marr and I published together, a paper titled "Directional Selectivity and Its Use in Early Visual Processing" [8]. The other paper was in a collaboration with Tommy Poggio, on a topic suggested by Tommy. He was intrigued by the power of sparse edges in the image to capture the image content, and we looked together on the

mathematical power of zero-crossings, used by Marr as precursors of image edges and contours, to capture image information [9]. David and I worked together on the directional selectivity paper extensively for a period of a few months. It was for me the first, and one of the only times, that I wrote a paper with someone in this mode, actually sitting together for long hours at a time, composing sentence after sentence, and discussing each paragraph we wrote. The experience was very intense and enjoyable. I think we both enjoyed it and were both exhausted by the time the paper was finished. With a British restraint, a footnote on the first page of the published paper read: "During the course of the publication of this paper it was learnt with regret of the death of Dr. David Courtenay Marr at the early age of 35."

As I studied different aspects of visual motion, I was continuously impressed by the complexity of the problem from a computational standpoint, and by the various mechanisms devised by biological systems to deal with the measurement and the use of motion information. Sophisticated mechanisms for extracting and utilizing visual motion are found even in simple animals, such as the frog and the housefly. As already mentioned, it appears that the visual system uses two complementary methods for extracting visual motion. One based directly on the local changes in light intensity (intensity-based methods). The other measures motion by the matching of features (such as edges, corners, blobs) across space and time (feature-matching schemes). My initial studies focused on feature-matching schemes, and their contribution to the perception of structure from motion. I later became more focused on intensity-based schemes, studying computationally how to measure visual motion based on image intensity changes and considering possible biological mechanisms for the task.

Computationally, intensity-based motion measurement is difficult, because the motion measurements are locally ambiguous, because of what I termed the "aperture problem" [10]. A spatial integration stage is required to resolve the measurement ambiguities. The integration is a difficult problem, which is still not entirely resolved. Because of integration problems, computational schemes for motion measurement tend to produce blurry and inaccurate velocity measurements at motion boundaries (e.g., between a moving object and its background). Ellen Hildreth, my first doctoral student, worked on the motion measurement problem for her excellent dissertation, published in a book form under the title *The Measurement of Visual Motion* as an Association for Computing Machinery Distinguished Dissertation [11].

Other studies of intensity-based motion measurement, focusing on models of biological mechanisms, were conducted in collaboration with my close friend Jacob (Kobi) Richter. Kobi has been a friend since my air force days. He was somewhat senior to me and already a well-known ace pilot when

we met. We spent time together training to become flight instructors and kept close connections since. Independently, Kobi's wife Judith and my wife Chana were friends even before I met Chana. Kobi was (and still is) a multitalented person, excelling in everything he puts his mind to. He remained in active air force service, but in parallel to his service, he completed a doctorate in physiology at Tel-Aviv University. In 1979, he took a three-year leave of absence from service, obtained a fellowship and research position at MIT, and came to Boston with Judith and their three children. At MIT he worked with me on computational modeling and with Peter Schiller on monkey V1 recordings. Kobi quickly became an accomplished programmer and modeler. In a number of publications we had together, modeling temporal aspects of retinal ganglion cells and V1 simple cells, Kobi provided the empirical data and participated on equal footing in the modeling work [12–14].

In addition to visual motion and some other modeling work, a topic that occupied me conceptually was the role of computational modeling in understanding brain and cognition. Related to this were attempts to explain to colleagues and fellow scientists what we were trying to do and in what ways computational theories and models could complement empirical studies. This was not always easy. A well-known professor from Yale invited me once to give a seminar in his department. He asked me, however, to "please leave out the computational and theoretical parts," because he did not think his colleagues would have any interest in these directions. As a part of this general direction I wrote a paper, which gave rise to a lively discussion, about Gibson's theory of direct perception [15].

James Gibson was a distinguished psychologist who made important contributions to the study of visual perception, including studies of the optical flow during locomotion, the role of texture gradients in the perception of surfaces, and introducing the interesting notion of "affordances." On the theoretical side, Gibson opposed any use of the terms "computation" and "representation" in the study of perception. In his view, information about the environment is "picked up directly" by the observer, without any need for further elaboration. Because this position is diametrically opposite to computational theories, it was of interest to analyze what "direct perception" means, what is exactly the argument, and what is it based on. The critique I wrote about direct perception was not about Gibson's contributions, but rather about his theoretical stance, which I find untenable as a scientific explanation.

In other work along the same goal of making connections between neuroscience and computer science, I wrote several general overviews, aimed at either computer scientists discussing relevant neuroscience aspects [10] or aimed at neuroscientists, discussing relevant computational aspects [6,16]. Over the years, there has been a continuous shift in the degree of understanding and collaboration between the biological, behavioral, and computational studies of vision. Many of the neuroscience graduate students I have

met in recent years at the Weizmann Institute, MIT, and other places are highly sophisticated in their computational background and skills, and they use these skills in their neuroscience work.

On the Applied Side: Orbotech

Toward the end of my term as a research scientist and before starting my faculty appointment, I spent six months as a visitor at the Weizmann Institute of Science in Israel. Chana and I felt that after spending seven years in Boston, we wanted to spend a significant time period with family and friends back in Israel. I thought initially of spending this time at my Alma mater, the Hebrew University in Jerusalem, by I got a phone call from an ex-teacher of mine in mathematics at the Hebrew University, who had moved in the meantime to the Weizmann Institute, urging me to come instead to the Weizmann, to the mathematics and computer science department. The Institute welcomed us warmly, arranging lovely accommodations and a nice office space.

During this stay, I gave in the department a class on computer vision. This was the first class on computer vision at the Weizmann, and in Israel in general, and it was well attended by both students and faculty members. Among the listeners in the class was a computer engineer, Zvi Lapidot, who found this new field highly exciting. We discussed topics in computer vision frequently after class and soon became good friends. Zvi was a part of a group in the computer science department that actually designed and built computers, building the hardware and writing the software to run it. This activity had begun already in the 1950s, with construction of the first computer in Israel, and one of the first large-scale computers in the world, called the Weizac. Building such computers during the 1950s was still a pioneering effort, and the Weizmann recruited for the task several scientists and engineers from the United States. The project had an advisory committee, which included John von Neumann, Albert Einstein, and Robert Oppenheimer as its members. Following the Weizac, two additional generations of computers were designed and built at the Weizmann, the Golem and its successor, the Golem B, which was a large-scale mainframe computer. Zvi was a software engineer in the Golem project, and because the Golem team was small and closely knit, he acquired deep knowledge of computer systems in general. As an engineer with a bright and curious mind, he was intrigued by the field of computer vision and the possible applications it could offer.

By the time my short term at the Weizmann was over, Zvi was excited by the possibility of moving into this new field. He took a leave from the Institute and found a job in a computer vision startup in the Boston area. He wanted to learn about the applied aspects of computer vision and to consider potential applications in the field. He eventually reached a conclusion that the manufacturing of printed circuit boards was at a critical stage

in terms of moving into automatic visual inspection. Many in the industry were skeptical about the possibility of automating the inspection process. Zvi raised the problem with me, and after a while, we reached the conclusion that new methods from computer vision could be of use for automating the inspection process. We added to the discussion two close friends: a computer engineer from the Weizmann named Ami Caspi who worked for years with Zvi, and my friend Kobi. Ami was a superb hardware engineer, and Kobi was superb at anything he became interested in. This was quite a unique group, and within a short time, we had the initial blueprints for a vision system for printed boards inspection. The plans moved quickly forward. When Zvi went back to Israel, a company was formally formed (under the name Orbot), funds where raised, and a small group of engineers was hired. Within two years, several systems were already operating in several printed circuit manufacturing facilities. After a merger with another Israeli company in the same area, the company name was changes to Orbotech. It was listed on Nasdaq in 1984 and expanded in a number of additional directions, including the visual inspection of flat-panel displays. In 2019, the company was acquired by KLA, a leader in process control for the semiconductor industry.

For me, working with the initial group of talented friends from different disciplines was highly stimulating and rewarding. Professionally, some of the problems involved in the high-speed analysis of complex patterns were interesting and challenging, in part as problems in vision and in part from more mathematical or algorithmic perspectives. As the company matured, my own involvement became small, but I stayed connected with the company until its final acquisition.

Faculty, between MIT and Weizmann

Following three years in a research position, I started my faculty appointment at MIT in 1981, at the rank of associate professor. The following year, 1982, was quite transformative for me and my family in three directions. First, Chana and I had our two sons Tomer and Yonatan. Second, my MIT position became tenured. Third, I also accepted a position in the computer science department at the Weizmann Institute of Science, in parallel with the MIT position. It was agreed that I will spend about half of my time at each institution, alternating every one or two years. The most dramatic transformation for us was of course becoming a family and starting to raise our children. I don't think that I fully anticipated how profound and gratifying this transformation would turn out to be. Many years have passed since then, and Tomer and Yonatan are now going through a similar experience, raising their own families. Yonatan chose to be an artist—a painter and sculptor. Tomer chose to be a scientist and is now a faculty member at Harvard's Psychology department. I think that both chose to pursue what they were always passionate about, which I hope will bring them lifelong satisfaction.

During this period my scientific research focus started to shift to increasingly higher levels of visual perception. My intellectual interests were in fact biased from the beginning of my studies of vision toward the higher levels of vision and the integration of vision and cognition, problems that include, for instance, object recognition, scene understanding, and aspects of visual awareness. Research in computer vision in its early days had very little to do with such topics, focusing instead on lower-level aspects of vision, such as edge detection, feature detection, texture separation, motion measurement and depth estimation. In my own work on motion, I was surprised by the complexity of the problems involved in motion perception and the sophistication of the computations performed by the brain in measuring visual motion and recovering three-dimensional structure from motion, as discussed earlier. One can certainly devote a rich lifelong scientific career to the study of vision at different levels; the move toward higher, more cognitive levels was not because I thought that lower levels of visual processing were already well understood, but I just following my personal scientific interests.

Visual Routines

One general topic I became curious about is the general nature of perception when performing complex visual tasks, such as scene understanding. Rather than a single application of a bottom-up process, behavioral evidence and computational considerations suggested the use of a sequential process, guided by top-down processes, to accomplish goals of interest. I addressed this general topic initially in the context of what I called "visual routines," first in computational studies [17], combined later with psychophysical studies.

To take a concrete example, consider the relation of inside versus outside, which is performed by the human perceptual system with intriguing efficiency. For instance, suppose that the visual input consists of a single closed curve and a small X figure. The visual task is to determine whether the X lies inside or outside the closed curve. If the shape is not highly convoluted, the "insideness" relation pops out immediately. But what is the visual process that allows us to reach this conclusion? A popular method that has been proposed for "insideness" in practical applications uses the so-called ray-intersection method. Starting from the x location, draw a ray in any direction to the edge of the image, and count the number of intersections along the ray. If the number is odd, the x lies inside, and otherwise it is outside the closed contour. This ray-intersection method works correctly only under limited conditions—for instance, it must be assumed that the curve is closed, and it must also be isolated, because the presence of additional contours can create extraneous intersections. It turns out that seeking a robust solution of the "insideness" problem is a challenging problem. For human vision, we also seek a biologically plausible computation. Biological plausibility is not a simple and well-defined criterion, but the ray-intersection

computation, for instance, appears to be an unlikely candidate. A method that is more computationally effective and fits better with human perception is based on spreading activation. Starting from a given point, the area around it in the internal representation is somehow activated. This activation spreads outward until a boundary is reached, but it is not allowed to cross the boundary. Depending on the starting point, either the inside or the outside of the curve, will be eventually covered by the spreading activation. This can provide a basis for separating inside from outside. Additional stages are still required to complete the procedure, and the additional stages will depend on the specific problem at hand. I will not consider here further "insideness" or visual routines in general, because the topic is too specialized for an overall review. Note, however, that the basic themes that played a central role in studying the perception of spatial relations by visual routines came up repeatedly in other studies of higher-level aspects of vision. Performing complex visual tasks, such as scene understanding, often employs an evolving sequential process, guided by top-down processes, to accomplish some goal of interest. In the study of visual routines, one of my goals was to combine the computational studies with empirical psychophysical and physiological data. Together with a wonderful psychophysics colleague, Pierre Jolicoeur, we have conducted a number of studies, focusing primarily on the task of boundary or curve tracing [18,19]. Some of this work was also followed subsequently by creative studies at the physiological level by Pieter Roelfsema and others [20,21].

Object Recognition

When I started to work on the broad and difficult problem of object recognition, I was asked by one of my colleagues whether I decided to stop working on vision. People worked at the time on aspects of pattern recognition, but studying natural object recognition was at its infancy, and at least some researchers considered object recognition as belonging more to higher level cognition than to vision research. Views regarding the relations between vision and cognition shifted over time to an almost opposite view. Object recognition became a central part of vision research, and current methods for object recognition perform very successfully, without employing more cognitive aspects, such as world knowledge, logical inference, or symbolic representations. The same methods that proved useful for object recognition, based on deep convolutional networks, are also being applied to a more cognitive and abstract level of using vision, such as scene understanding or visual question-answering. My own view is that application of these methods, with only minor alteration, to the higher and more abstract levels, is going to run into difficulties. I have worked on object recognition for a long time, and I think that my personal history during this period reflects much of the general evolution of the field during this period. I will not get into

technical details, but I will describe briefly some of the main attempts and conceptual shifts in object recognition and higher-level vision.

When I started this journey, the dominant approach to object recognition was based on three-dimensional representations and internal manipulations of these representations. Using three-dimensional representations to support recognition makes sense: I can recognize the object "car" from a large range of viewing directions, despite the large change in the car's image as the viewing direction changes from a frontal, to side, or back views. Influential models using this approach were Marr's generalized cylinders representation [22] and Biederman's theory of recognition by components [23].

In both theories, objects were represented using a collection of three-dimensional primitive shapes. My own initial work on object recognition was also based on three-dimensional representation but from a different perspective. I describe next some of the work and how it developed over time. The main objective is not to discuss details of object recognition, but to recount my experience in exploring a complex and little understood research area, until reaching a conclusion that a basic shift is needed. A related theme is how the research was guided by a combination of modeling results with empirical neuroscience evidence.

The basic motivation for using three-dimensional models for visual recognition was to achieve view invariance. In the components-based approach, invariance was achieved using the fact that the object components and their arrangement remained invariant to changes in viewing directions. I thought that these representations were too qualitative for recognition, and instead of relying on a vocabulary of shape primitives, I proposed that view-invariant recognition could be obtained by alignment [24].

The initial work along this line was conducted with my student Dan Huttenlocher, now the dean of the MIT College of Computing. Dan was a brilliant and lively student who soon became also a colleague. During his doctoral work, he spend some time with me at the Weizmann in Israel. I asked him many years later if he remembered any of the Hebrew he acquired during this period in Israel. The expression he remembered best was "nine percent" (in Hebrew). It turned out that he found a particular cheese that was available in grocery stores across Israel, which was just called "9 percept cheese"; he loved this cheese, and the useful term was imprinted in his memory.

The work with Dan was published in the proceedings of what can be seen in retrospect as somewhat of a historic event—the first meeting of the International Conference on Computer Vision (ICCV), which later became a major meeting in computer vision with many thousands of participants, a broad range of workshops, and a commercial exhibition. The central idea was that to recognize an object, an internal object model was first aligned with the image using a small number of model and image features. We used for this alignment results from the structure-from-motion work. For example, by identifying three corresponding feature points in the model and the

image, it becomes possible to compensate changes in position, scale, and spatial orientation between the model and the image, and then to compare the image directly with the aligned object projection.

The method was efficient and highly accurate, but it had drawbacks as a candidate model for natural object recognition. For example, it could recognize well a specific object, such as a specific car model, but had difficulties dealing with differences between individual objects that belonged to the same class, such as the general "car" class. We explored over a number of years some modifications and extensions. For example, aligning not only object contours but also more abstract shape descriptions, or using multiple local alignments rather than a single global one [25], I was gradually led to the conclusion that, in general, human recognition of natural objects does not rely on the use of three-dimensional object models. This conclusion was based primarily on computational considerations, namely the difficulty that such models have dealing with general classification and with highly flexible objects.

About 10 years after our initial work on alignment, evidence from neuroscience also started to supply evidence against the involvement of internal three-dimensional manipulation in object recognition. Psychophysical work by Shepard [26] showed that humans can perform some form of internal manipulation to bring three-dimensional alignment, a capacity known as "mental rotation." Such studies may seem to support the possibility of three-dimensional alignment in recognition. There are, however, good reasons to doubt the connection between mental rotation and object recognition. First, mental rotation is a slow process, unlikely to support fast object recognition. Second, brain imaging studies, starting in 1996 [27], showed that brain activation during mental rotation is mainly in areas not directly involved in natural object recognition.

The brain has an ability to perform some form of objects rotation and alignment, but this capacity appears to be associated more with tasks related to motor planning, such as inserting one object into another. The internal object manipulation can still be useful for recognizing objects from unfamiliar views. Humans can recognize objects from completely novel views, and their ability in this task surpasses the ability of the leading computer vision models available today [28]. This ability may be attributed at least in part to an internal alignment process, but I became convinced that general object recognition by humans relies on other methods, still to be explored.

I find that the history of using three-dimensional object models and alignment methods provides a nice example of useful interactions between computational modeling and direct neuroscientific evidence. The computational investigations by themselves provided evidence regarding the limitations of alignment methods for general object recognition. The neuroscientific data supplied more direct evidence against alignment methods, and more generally influenced me and others in our thinking about possible mechanisms and models of object recognition.

The work on high-level vision up to these early recognition models included some interesting topics not described here, in particular image saliency, done with my long-time friend Christof Koch [29], and an early model of combining bottom-up with top-down processing in the visual cortex [30]. These topics are summarized in the book *High-Level Vision: Object Recognition and Visual Cognition* [31].

Recognition by Image Patches

The shift from three-dimensional modeling to a different approach to recognition was a pretty extreme one. Instead of focusing on differences caused by changes in viewing directions, it appeared to me to be more important to focus on dealing with the difference in appearance caused by differences between objects belonging to different classes. Humans are better at recognizing objects at the class level (e.g., a car or motorcycle) than recognizing individual cars and can also easily tolerate complex nonrigid transformations. They also do not need the entire object; a partial view, sometimes even a small part of an object, may be sufficient. These and related considerations led me to the direction of basing the recognition of a class of objects based on selected image patches that are specific to a class of objects, taken directly from example views of objects in the same class. That is, the shape fragments used to represent faces, for instance, would be different from the shape fragments used to represent cars or letters in the alphabet. These fragments can be used as a set of common building blocks to represent, by different combinations of the fragments, different objects belonging to the class. The patch-based representation appealed to me for the purpose of object classification based on general observations, but it was also supported by empirical tests, showing that objects from the same class shared many similar local patches and that local patches by themselves are highly informative for the class identity. The problem of recognition across different viewing directions became secondary. It can be approached by the class-based fragments approach by using for each class a number of related models, from different viewing directions [32,33].

We introduced the patch, or fragment-based approach in the late 1990s, along with a general method for selecting the most informative patches of a class for the purpose of recognition. The selection of representative class fragments was based on a natural notion from information theory called mutual information. Without getting to the technical details, the notion is quite intuitive. Suppose you know that an object was placed in one of eight boxes, but you do not know which box. Suppose also that you get a hint—that the box containing the object is red, and four of the eight boxes are red. The hint about the color reduces the uncertainty of the choice from eight to four possibilities. Measured in bits, the uncertainty was reduced from three bits (the number 8) to two (the number 4). The saving of one bit was the

mutual information supplied by the color hint. In a similar manner, one can measure how the presence of a particular patch in an object's image reduces the uncertainty about the object class and then can select, in an automatic procedure, class patches that are the most informative for classification. An interesting finding with some biological and computational implication was that the most informative fragments for classification were class fragments of intermediate complexity, compared with smaller local features or complex patches that covered most of the object [34].

The processes of selecting informative object features can proceed in a hierarchical manner, using small informative patches to detect larger patches. One gets from this object representations in terms of a hierarchy of informative patches, which often correspond to semantically meaningful parts and subparts. An exciting possibility offered by such hierarchical object representations was that recognition may take place at all levels simultaneously, recognizing not only the object, say a face, but also its parts and small subparts, such as eyes as well as the eyelids, lashes, and pupil. In a model that brought together much of our work on fragment-based recognition, we obtained this kind of full recognition. An interesting finding was that such a full object interpretation, identifying and localizing parts at all levels, required a combination of bottom-up with top-down processing. The interpretation used a single cycle of this kind, a bottom-up pass that identified the object class, followed by a top-down pass, which used the recognition of one level to identify parts at a lower level. The top-down pass used the context at a higher level to disambiguate parts at a lower level, which could not be recognized using the bottom-up recognition [35].

Current deep network models for object classification also create a hierarchy of increasingly complex features, which lead to excellent classification results. In their current state, however, these models do not combine object classification with the recognition of parts and subparts at multiple levels. It would be interesting if future work could combine deep net models for object classification with a parts recognition stage, perhaps using a similar approach to the bottom-up, top-down cycle discussed earlier.

As in other areas, I tried to combine the computational modeling with empirical behavioral and physiological data to examine the applicability of the models to human vision. In one study, [36], we collaborated with Shlomo Bentin, a highly creative neuroscientist from the Hebrew University, to test the role of informative features in human vision using electroencephalogram (EEG) and behavioral tests. We showed subjects images of different fragments, for which we calculated the mutual information they provided for classification, and measured the subjects' classification accuracy and EEG responses. We found that categorization performance correlated with the measured mutual information level as well as with the amplitude of a posterior temporal potential, peaking around 270 milliseconds. Another study, in collaboration with Rafi Malach [37], obtained similar results using

functional magnetic resonance imaging (fMRI): it was possible to use the objective measure of mutual information between an image patch and a class to predict the level of fMRI activation in human object areas. For me, there was some scientific satisfaction in the ability to take an image patch and make a prediction about classification accuracy and the brain response it would induce when shown to a human observer.

The ability to recognize objects based on local patches and other forms of reduced information had a revival some 10 years later, when we examined what we called "atoms of recognition" in human vision [38]. This was another attempt to identify the visual features and representations used by the brain to recognize objects. The studies of informative class fragments described earlier identified object fragments for classification. An object in the image would typically be covered by multiple fragments, which together, by summing up their individual contributions, would lead to the object's classification. In studying "atoms of recognition," we used a large-scale psychophysical study to systematically search for the minimal images that are sufficient for reliable recognition. For each class of image we tested, such as a horse, bird, bicycles, and flies, we found multiple "minimal images" that were sufficient on their own for a reliable recognition of the object in the image. An interesting finding was that at the level of minimal recognizable images, a minute change in the image can have a drastic effect on recognition, thus identifying features that are critical for the classification task. Simulations then showed that existing recognition models could not explain this sensitivity to minor changes, and, more generally, they did not learn to recognize minimal images at a human level. The role of the critical features for recognition is revealed uniquely at the minimal level, where the contribution of each feature is essential.

In a study led by Marlene Behrmann from Carnegie Mellon University [39], we used the set of minimal images we identified to probe the responses of class-specific cortical regions in humans using fMRI. In particular, we compared the minimal images with subminimal images—these were slightly reduced versions of the minimal images, which were only slightly different but essentially unrecognizable. As expected, because of their similarity, in early visual cortex, the minimal and subminimal produced similar responses. Higher-level, class-specific regions, however, exhibited greater activation for minimal images compared with their subminimal counterparts. Moreover, minimal images from each category elicited enhanced activation in corresponding category-selective regions—for example, the parahippocampal place area showed selectivity in its response already at the minimal image level.

Object classification is often considered today to be a solved problem, because deep network models excel in this task. This is a remarkable achievement both in practice as well as in our understanding of biological mechanisms of classification. Human classification performance, however, is

still qualitatively better than models in various aspects (e.g., unsupervised learning, out-of-distribution examples, adversarial examples, unfamiliar viewing directions, or minimal images), and additional studies and comparisons with human vision are still required to fully understand the object classification task.

Digital Baby

Learning to recognize objects is an interesting and challenging problem in understanding vision, but it is only a part of a much larger problem, of using vision to learn about the world. This learning process starts in infancy, when much of the knowledge about the world emerges from the combination of innate mechanisms and visual experience. I have been, and still am, fascinated by this larger problem, and for a number of years, I have been engaged in a research project I called the "digital baby," exploring basic issues within this large domain. The ultimate goal of such a project is to develop a digital baby model that, through perception and interaction with the world, will develop on its own representations of complex concepts, which will allow it to understand the world around it, in terms of objects, categories, events, agents, actions, goals, social interactions, and the like.

This emergence of understanding is a major challenge in the study of vision, cognition, and the brain. Current computational theories dealing with the acquisition of knowledge about the world through visual perception still cannot cope with this major challenge. They have made an impressive and substantial progress over the past decade, but as far as I can judge, current methods are unable to acquire spontaneously and deal effectively with natural cognitive concepts, which depend not only on statistical regularities in the sensory input but also on their significance and meaning to the observer.

A major part of the problem in my view is that current computational approaches to visual learning are too empiricist in nature. As has been shown by a rich body of developmental studies, the human cognitive system is equipped through evolution with basic innate structures that facilitate the acquisition of meaningful concepts and categories. These are used to obtain a true understanding of the world, which goes beyond correlations and statistical regularities. In 2010, the Cognitive Science Society published a brochure titled "Outstanding Questions in Cognitive Science." In it, I suggested that a basic open question in cognitive science is a theory of "computational nativism"—a computational theory of cognitively and biologically plausible innate structures and how they guide the cognitive system along specific paths through its acquisition of knowledge to continuously acquire meaningful concepts and useful representations.

I will describe briefly a couple of examples from the digital baby studies that illustrate the power of plausible innate structures to guide the spontaneous acquisition of meaningful concepts. The goal in describing these

examples is not to present our work in detail, but rather to discuss major issues in perception and cognition that I have been interested in for a long time and to describe briefly my general views on some of these issues.

In the first example, we developed a model that tackled two notable problems in which the gap between computational difficulty and infant learning is particularly striking: learning to recognize hands and learning to recognize gaze direction [40]. Hands are known in computer vision to be objects that are particularly difficult to recognize; because they are highly articulated, they can appear in a large variety of different shapes. In addition, in many cases, their size in the image may be small, and their recognition is obtained not based on their own appearance, but mainly by the context of other body parts. And yet, hands are one of the first objects to be recognized reliably by infants, starting already around the age of three months. Direction of gaze, and the target of someone's gaze, are similarly difficult, but again, gaze direction is learned early in infants' development. Unlike computational models, this learning is obtained in an unsupervised manner, that is, without any help or guidance by a teacher or by annotated images. Infants just look at the world around them and spontaneously develop the ability to recognize hands and direction of gaze.

The model we proposed shows similar capabilities: it is shown a stream of natural videos, and it learns without any supervision to detect human hands by appearance and by context, as well as by direction of gaze, in complex natural scenes. How can such complex learning take place spontaneously by "mere looking"? We propose that the learning process is guided by an innate mechanism—that is, the detection of what we called "mover" events in dynamic images. Such an event is defined by a moving image region causing a stationary region to move or change after contact. The proposal is based on two types of evidence. The first is evidence regarding infants' sensitivity to special types of motion, similar to mover events. From an early age, infants are sensitive to visual motion, and they use motion to separate moving regions from a stationary background and to track a moving region. They are also sensitive to specific types of motion, similar to our proposed "movers," including so-called launching events, as well as to active motion (causing other objects to move) or to self-propelled motion. The second part of the evidence comes from our computational studies, which demonstrate that having the capacity to detect specific motion events in dynamic images can naturally lead to the automatic acquisition of increasingly complex concepts and capabilities, which do not emerge without domain-specific biases. The detection of mover events on their own are not sufficient for hand recognition: they do not identify the hand as an object, they are not sufficiently reliable, and they detect hands and some hand configurations but not others. However, reliable and general hand detection can evolve from a combination of the teaching signal provided to the mover events, with existing methods for supervised learning of object recognition, using mover events

as providing the supervision. After exposure to video sequences containing people performing everyday actions, and without external supervision, the model develops the capacity to locate hands in complex configurations by their appearance and by surrounding context.

This combination of internal structures with learning appears to me to be a powerful and general combination. Infants learn to recognize hands in a spontaneous manner early in their development, despite the fact that hand recognition is difficult and does not emerge in learning models by just looking at the world. This is obtained in the model not by using innate hand detectors, but by using simpler "proto concepts" for hands, which guide the learning process along a trajectory that leads to hand recognition. Note that during the learning trajectory leading to hand recognition, hands become naturally associated with some of their meaningful properties, such as hands being the mover of objects, and they are used for object manipulation and actions. Developmental studies show that young infants indeed expect a human hand, and not an inanimate object, to be the primary cause of an inanimate object's motion [41].

The model for learning to recognize hands was also extended using similar principles to the learning of extracting direction of gaze. I will not get into any details of gaze detection, except to mention what was the teaching signal used in this case to guide the learning. Based on empirical observations, we made the assumption that when people make a contact with an object, to grasp and manipulate it, they almost invariably look directly at the object. They can subsequently move the object around without looking at it, but at the point of making the initial contact, the contact location is a reliable cue to the direction of gaze. Again, the target concept, gaze direction in this case, is difficult to learn by just looking. It appears early and spontaneously not based on an innate structure in our visual system that serves for gaze detection, but rather by the use of a proto concept in a learning process that leads to the reliable acquisition of the final target capacity. This learning process also naturally links the gaze direction and the gaze target with the goal and attention of the looker. As in the hands example, internal structures guide the system to acquire meaningful concepts, which are significant to the observer but statistically inconspicuous in the sensory input.

We recently had an opportunity, in a study led by Ehud Zohary, to test the implications of the gaze-recovery model with a special population of patients in Ethiopia who had early bilateral congenital cataract, diagnosed and treated at late childhood [42]. This sight restoration provided a unique opportunity to directly address basic issues on the roles of "nature" and "nurture" in development, as it caused a selective perturbation to the natural process, eliminating some gaze-direction cues while leaving others still available. Of particular interest was the ability of the group of patients who did not have sufficient visual acuity before surgery to recover fine details (e.g., the position of the irises) needed to recover the direction of gaze. Following

surgery, the patients' visual acuity typically improved substantially, allowing for fine discrimination of eye position. Yet, the recovered patients failed to show eye-gaze following behavior, and they also fixated less on the eyes of a viewed face, two reflexive behaviors seen in controls, even years after vision was restored. This is a natural expectation of the model, assuming that the specialized internal guiding signals that normally enable the learning of eye-based gaze direction are no longer available at a later stage, when the correcting surgery was performed. As expected, their gaze-following behavior based on head rather than eye direction was typically normal, which was consistent with the available visual acuity before surgery.

Another digital baby example of innate structures guiding learning is the surprising acquisition of concepts related to containment and containers [43]. Containment is one of the earliest spatial relations to be learned, starting around three months of age, and preceding other common relations (e.g., support, in between). Computationally, containment often depends on subtle cues in the image, and it is unclear how this relation is acquired so early in development and without supervision. The model we proposed can explain infants' capacity of learning containment and related concepts by "just looking," along with their empirical development trajectory. Learning occurs in the model fast and without guidance, relying only on perceptual processes that are present in the first months of life. As in the learning of hands and gaze direction, instead of labeled training examples, the model provides its own internal supervision to guide the learning process. We showed how the detection of so-called paradoxical occlusion provides natural internal supervision, which guides the system to gradually acquire a range of useful containment-related concepts. During early development, even earlier than acquiring containment, infants develop sensitivity to occlusion relations. The digital baby model includes a similar capacity—that is, to use visual motion to learn occlusion relations along with dynamic and static object segregation [44]. We showed how the notion of containment can naturally arise from a simple combination of occlusion relations, which I will not describe here in more detail. The model works well, and it acquires different aspects of containment in an order that is similar to the temporal learning order by infants [45].

Overall, these examples of the digital baby model illustrate how natural complex concepts can be acquired visually without requiring elaborate external supervision. The results show how combining perceptual learning of the kind obtained by deep learning with brain-like innate structures may guide models toward human-like learning [46].

In my view, future models of human vision, as well as intelligent AI vision systems, will be closer to the general structure suggested by the digital baby model than current models, which utilize primarily unguided bottom-up processing. Unlike networks that start from a tabula rasa state and that learn using extensive supervision and huge amounts of annotated data,

human-like models will incorporate rich innate structures that guide the learning along useful learning trajectories with limited supervision. Guided by innate structures, the learning process will construct a rich conceptual system for understanding the world around us.

Top-Down Processing, Linking Perception and Cognition

Another major topic of interest to me for many years has been the role of top-down processing in vision, and the integration of vision and cognition in scene perception. Anatomically, we know that the visual system combines closely interacting bottom-up and top-down processing streams. The bottom-up pathways go from early to higher level areas, including more abstract and cognitive brain regions, and the top-down pathways proceed in the opposite direction, from high to lower regions. The cortical circuitry suggests that the top-down stream plays a major role in vision and in cortical computations in general, because it is often at least as massive as the bottom-up stream. The exact function of the top-down projections is still unclear, and given its importance, this is obviously a major open question, which may shed light on basic aspects of brain processing and on the integration of vision with higher-level cognitive functions. As noted, current deep network models, which excel in visual object recognition, perform the task relying almost exclusively on bottom-up processing. If the complex task of visual recognition can be performed well without requiring a top-down component, what can be the core contributions of the top-down pathway?

In my own work, starting around 2000, the initial direction I followed was to explore the use of the top-down stream for the specific task of object segmentation, also called figure-ground separation. When we recognize an object in the image (e.g., a horse), we can also identify the exact image region that contains the horse. Most approaches to segmentation at the time assumed that separating an object from its background is performed based on image properties, such as uniformity of color and texture, or the smoothness and continuity of the object's bounding contours in the image. The problem is that natural objects can be highly nonuniform, with multiple colors and textures and jagged boundaries, and can still be well segmented and separated from their background. It is therefore plausible that once we become familiar with a given type of object (e.g., what a horse looks like), we have some form of model of the object in higher-level visual areas, which can be used for guiding the segmentation process at lower levels. In work with E. Bornstein, we proposed a model that showed for the first time how an object model can be learned from a set of images, and then be used to guide a top-down segmentation process [47,48].

The most flexible and accurate segmentation model was obtained by combining bottom-up and top-down processes [49,50]. The object-model can combine object parts that are highly variable in terms of image properties. In contrast, low-level image properties, such as the exact location of edges,

are highly informative for delineating precisely the object's boundaries. It is not surprising, therefore, that in current deep network models of vision, the segmentation task is usually performed by a top-down network, combined with the bottom-up one. Experimental evidence from the neuroscience and psychophysics of perception has provided detailed evidence supporting the general view that bottom-up analysis of the retinal image is combined with top-down processing that include stored memories and guiding signals to obtain accurate and useful perceptual interpretation [51]

In subsequent work, we proposed another task in which bottom-up and top-down processing complement each other well, which is the task of recognizing and segmenting object parts and subparts at multiple levels (e.g., a person, arm, hand, finger) and even finer details (e.g., a fingernail), if the resolution is sufficient. This work was mentioned earlier, in the context of performing patch-based image interpretation, using a bottom-up, top-down cycle. The use of a top-down component is again natural for the task, because the identification of a given part supplies context that can be used to disambiguate parts at a lower level, which could not be recognized using the bottom-up recognition alone [35].

In both of these models, object segmentation and multilevel parts recognition, the connections between the bottom-up and top-down streams go in one direction only, from the bottom-up to the top-down stream. In more recent work, we developed a more general framework, in which the connections are bidirectional, and the goal is more ambitious: to obtain scene interpretation by an iterative bottom-up, top-down processing [52].

Scene Interpretation

Full scene perception is still a major open problem in understanding and modeling human vision. This is a complex and challenging task, because it requires the extraction and representation of scene components, such as objects and their parts, people, and places, along with their individual properties, as well as relations and interactions between them. Computational models of scene perception, also called "scene interpretation" models, have focused on training network models to extract a structural representation of the entire scene, with all its components, and the relationships between the components. In contrast, humans' scene perception focuses on selected structures in the scene, starting with a limited interpretation (the scene "gist") and evolving sequentially, in a goal-directed manner. It seems to me that extracting a full scene description is both infeasible and unnecessary, and therefore the selective interpretation produced by human perception offers an attractive alternative to current models. However, it also raises new difficulties, of extracting selected scene structures of interest in a goal-guided manner and applying a sequential process that depends on both the image content and the current observer's goal.

In our work, we proposed to obtain human-like goal-directed scene interpretation, using an iterative bottom-up, top-down processing, in a "counter-streams" structure motivated by cortical circuitry. In this iterative process, each cycle is composed of a bottom-up and a top-down part. The bottom-up part is a standard visual stream, which delivers useful visual representations to the higher, more cognitive components of the perceptual system. At this higher level, the extracted visual representation can be augmented with relevant nonvisual information. The higher-level areas also provide input to the top-down stream, which selects the relevant information to extract next. The top-down stream then guides the visual stream to extract the selected information in the next cycle. This guidance is obtained using the cross-stream connections, from the top-down to the bottom-up streams: through these connections, the next bottom-up pass takes place in the context of the top-down representation, and consequently, the information extracted by the bottom-up visual stream will depend on both the image and the top-down instruction provided by the higher levels.

I will not discuss the scene perception process in more detail, but only note that the interactions between the bottom-up and top-down streams seem to me to be a key factor in the continuous integration of vision and cognition in the perceptual process. As previously described, at each cycle of the scene interpretation process, visual and nonvisual information are combined in the extracted representations, and the combined representation then determines the top-down instruction for the next cycle. The results are, first, that the visual process can be guided effectively toward scene structures of interest to the viewer, and second, the perceptual outcome will naturally combine visual with nonvisual, more cognitive sources of information. This integration seems to be a key aspect for modeling human scene perception. It may also be critical to advanced AI vision systems, which currently excel at extracting visual information but seem to have difficulties with combining vision and cognition in the perception of visual scenes. The area of combining vision and cognition continues to be for me the main focus of continuing research, because it gets to the heart of the problem of using vision to understand the world.

Let me end with some comments on the role of my subarea of lifelong research within the huge and diverse field of the human brain and its functions, and on the way this subarea is currently going. The approach taken in this research area may be described as "functional computational neuroscience," in which functional means that the models under investigation combine biological data, together with the capacity of the model to successfully perform relevant brain functions. For example, a model dealing with binocular vision in the brain, will consider the relevant anatomy and physiology of the visual cortex, and at the same time, it will also be able to perform the functional task of recovering three-dimensional scene structure from a corresponding pair of retinal images.

The power of functional models to help the modeling and understanding of brain structures recently received compelling support from the use of deep network learning, in particular in modeling visual processing in the primate visual stream. For example, a systematic study [53] compared a range of deep network models in terms of their performance in object recognition tasks, along with their precision in predicting spiking patterns in individual units along the visual pathway. The comparisons showed that, within a broad range, optimizing the model in terms of its performance in the visual task also tends to increase its ability to predict spiking responses to complex naturalistic images, at both the single site and population levels in different cortical areas. As a result of such relations between function and structure, attempts to develop functional models of the brain based at least in part on deep network learning became in recent years a highly active area. A large-scale example is an integrative effort, supported by the so-called brain-score project, which is a platform for evaluating models on how well they predict neural and behavioral brain measurements in different domains [54] (see also https://github.com/brain-score).

Although deep network models are making significant contributions to computational neuroscience, an open and intriguing question is whether this kind of modeling is sufficient to produce the perceptual and cognitive capacities of the human brain. Deep learning models are based on end-to-end learning using vast training datasets. The learning process is from scratch, in the sense that the initial model, before learning, is often a uniform structure, with minimal or without built-in modularity, and without specialized mechanisms that help both the development and the functioning of the mature system. In contrast, biological brains incorporate structures that have developed through evolution over a long period, which help both the development and the final functioning of the brain.

It might be argued that if we consider both evolution and individual learning together as a long trial-and-error learning, then an extended end-to-end learning process, combining evolution and individual learning, must be sufficient for acquiring all brain functions, including perception and cognition. However, the feasibility of such an extended learning approach with current deep learning technologies is still an intriguing open question.

Despite many impressive achievements in visual and cognitive tasks by current network models and training methods, fundamental challenges persist. One limitation, compared with human learning, is the extensive and continuously growing use of huge, supervised datasets. Another limitation is the limited capacity of current schemes to generalize well beyond the distribution represented by the training examples, and there are others. Such limitations may be related to the lack of innate mechanisms and built-in principles, which are incorporated in the brain but not in network models (as discussed in the digital baby examples). If these components are indeed essential, how might we discover and integrate them into functional

models? One possibility is that similar components might naturally emerge using existing methods, with some improvements, but relying primarily on an increase in model size and the available training data. Alternatively, upscaling may not suffice, and the missing components might be discovered either by novel computational methods (e.g., some form of efficient evolutionary-like search) or by adopting methods that might be unraveled by ongoing and future studies of the human brain and human cognition. These are fundamental and important open questions, which will become clearer with future advances in both human and machine intelligence, including, in particular, functional computational neuroscience.

Bibliography

[1] J.N.J. McManus, W. Li, C.D. Gilbert, Adaptive shape processing in primary visual cortex, *Proc. Natl. Acad. Sci. U.S.A.* 108 (2011). https://doi.org/10.1073/pnas.1105855108.

[2] A. Sha'ashua, S. Ullman, Structural saliency: The detection of globally salient structures using a locally connected network, in *IEEE Proceedings* (1988). https://doi.org/10.1109/ccv.1988.590008.

[3] A. Shashua, S. Ullman, Grouping contours by iterated pairing network, *Adv. Neural Inf. Process. Syst.* (1990).

[4] H.C. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, *Nature* 293 (1981). https://doi.org/10.1038/293133a0.

[5] S. Ullman, Maximizing rigidity: the incremental recovery of 3-D structure from rigid and nonrigid motion, *Perception* 13 (1984). https://doi.org/10.1068/p130255.

[6] S. Ullman, The measurement of visual motion. Computational considerations and some neurophysiological implications, *Trends Neurosci.* 6 (1983). https://doi.org/10.1016/0166-2236(83)90081-4.

[7] S. Ullman, *The Interpretation of Visual Motion*, MIT Press (1979). https://doi.org/10.7551/mitpress/3877.001.0001.

[8] D. Marr, S. Ullman, Directional selectivity and its use in early visual processing, *Proc. R. Soc. London Biol. Sci.* 211 (1981). https://doi.org/10.1098/rspb.1981.0001.

[9] D. Marr, S. Ullman, T. Poggio, Bandpass channels, zero-crossings, and early visual information processing, *J. Opt. Soc. Am.* 69 (1979). https://doi.org/10.1364/JOSA.69.000914.

[10] S. Ullman, Analysis of visual motion by biological and computer systems, *Computer* 14 (1981). https://doi.org/10.1109/C-M.1981.220564.

[11] E.C. Hildreth, *The Measurement of Visual Motion*, MIT Press (1984).

[12] J. Richter, S. Ullman, A model for the temporal organization of X- and Y-type receptive fields in the primate retina, *Biol. Cybern.* 43 (1982). https://doi.org/10.1007/BF00336975.

[13] J. Richter, S. Ullman, Non-linearities in cortical simple cells and the possible detection of zero crossings, *Biol. Cybern.* 53 (1986). https://doi.org/10.1007/BF00342887.

[14] J. Richter, S. Ullman, Are non-directional simple cells constructed from directional subunits?, *Biol. Cybern.* 54 (1986). https://doi.org/10.1007/BF00318427.
[15] S. Ullman, Against direct perception, *Behav. Brain Sci.* 3 (1980). https://doi.org/10.1017/S0140525X0000546X.
[16] S. Ullman, Artificial intelligence and the brain: Computational studies of the visual system, *Annu. Rev. Neurosci.* 9 (1986). https://doi.org/10.1146/annurev.neuro.9.1.1.
[17] S. Ullman, Visual routines, *Cognition* (1984). https://doi.org/10.1016/0010-0277 (84)90023-4.
[18] P. Jolicoeur, S. Ullman, M. Mackay, Visual curve tracing properties, *J. Exp. Psychol. Hum. Percept. Perform.* 17 (1991). https://doi.org/10.1037/0096-1523.17.4.997.
[19] P. Jolicoeur, S. Ullman, M. Mackay, Curve tracing: A possible basic operation in the perception of spatial relations, *Mem. Cognit.* 14 (1986). https://doi.org/10.3758/BF03198373.
[20] P.R. Roelfsema, V.A.F. Lamme, H. Spekreijse, Object-based attention in the primary visual cortex of the macaque monkey, *Nature* 395 (1998). https://doi.org/10.1038/26475.
[21] P.R. Roelfsema, P.S. Khaya, H. Spekreijse, Subtask sequencing in the primary visual cortex, *Proc. Natl. Acad. Sci. U.S.A.* 100 (2003). https://doi.org/10.1073/pnas.0431051100.
[22] D. Marr, H.K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes., *Proc. R. Soc. London Biol. Sci.* 200 (1978). https://doi.org/10.1098/rspb.1978.0020.
[23] I. Biederman, Recognition-by-components: a theory of human image understanding, *Psychol. Rev.* 94 (1987). https://doi.org/10.1037/0033-295X.94.2.115.
[24] D.P. Huttenlocher, S. Ullman, Object recognition using alignment, in *Proceedings of the 1st International Conference on Computer Vision* (1987).
[25] S. Ullman, Aligning pictorial descriptions: An approach to object recognition, *Cognition* 32 (1989). https://doi.org/10.1016/0010-0277(89)90036-X.
[26] 171 (1971). https://doi.org/10.1126/science.171.3972.701.
[27] M.S. Cohen, S.M. Kosslyn, H.C. Breiter, G.J. Digirolamo, W.L. Thompson, A.K. Anderson, S.Y. Bookheimer, B.R. Rosen, J.W. Belliveau, Changes in cortical activity during mental rotation: A mapping study using functional MRI, *Brain* 119 (1996). https://doi.org/10.1093/brain/119.1.89.
[28] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, B. Katz, ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models, in *Advances in Neural Information Processing Systems* (2019).
[29] S. Ullman, C. Koch, Shifts in selective visual-attention: Towards the underlying neural circuitry, *Hum. Neurobiol.* 4 (1985). https://10.1007/978-94-009-3833-5_5.
[30] S. Ullman, Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex, *Cereb. Cortex.* 5 (1995). https://doi.org/10.1093/cercor/5.1.1.
[31] S. Ullman, *High-Level Vision: Object Recognition and Visual Cognition*, MIT Press (1996). https://doi.org/10.7551/mitpress/3496.001.0001.

[32] E. Sali, S. Ullman, Combining class-specific fragments for object classification, in *Proceedings of the British Machine Vision Conference* (1999). https://doi.org/10.5244/c.13.21.

[33] E. Bart, E. Byvatov, S. Ullman, View-invariant recognition using corresponding object fragments, in *Lect. Notes Comput. Sci.* 3022 (2004). https://doi.org/10.1007/978-3-540-24671-8_12.

[34] S. Ullman, M. Vidal-Naquet, E. Sali, Visual features of intermediate complexity and their use in classification, *Nat. Neurosci.* 5 (2002). https://doi.org/10.1038/nn870.

[35] B. Epshtein, I. Lifshitz, S. Ullman, Image interpretation by a single bottom-up top-down cycle, *Proc. Natl. Acad. Sci. U.S.A.* 105 (2008). https://doi.org/10.1073/pnas.0800968105.

[36] A. Harel, S. Ullman, B. Epshtein, S. Bentin, Mutual information of image fragments predicts categorization in humans: Electrophysiological and behavioral evidence, *Vision Res.* 47 (2007). https://doi.org/10.1016/j.visres.2007.04.004.

[37] Y. Lerner, B. Epshtein, S. Ullman, R. Malach, Class information predicts activation by object fragments in human object areas, *J. Cogn. Neurosci.* 20 (2008). https://doi.org/10.1162/jocn.2008.20082.

[38] S. Ullman, L. Assif, E. Fetaya, D. Harari, Atoms of recognition in human and computer vision, *Proc. Natl. Acad. Sci. U.S.A.* 113 (2016). https://doi.org/10.1073/pnas.1513198113.

[39] Y. Holzinger, S. Ullman, D. Harari, M. Behrmann, G. Avidan, Minimal recognizable configurations elicit category-selective responses in higher order visual cortex, *J. Cogn. Neurosci.* 31 (2018). https://doi.org/10.1162/jocn_a_01420.

[40] S. Ullman, D. Harari, N. Dorfman, From simple innate biases to complex visual concepts, *Proc. Natl. Acad. Sci. U.S.A.* 109 (2012). https://doi.org/10.1073/pnas.1207690109.

[41] R. Saxe, J.B. Tenenbaum, S. Carey, Secret agents: Inferences about hidden causes by 10- and 12-month-old infants, *Psychol. Sci.* 16 (2005). https://doi.org/10.1111/j.1467-9280.2005.01649.x.

[42] E. Zohary, D. Harari, S. Ullman, I. Ben-Zion, R. Doron, S. Attias, Y. Porat, A.Y. Sklar, A. McKyton, Gaze following requires early visual experience, *Proc. Natl. Acad. Sci. U.S.A.* 119 (2022). https://doi.org/10.1073/PNAS.2117184119/-/DCSUPPLEMENTAL.

[43] S. Ullman, N. Dorfman, D. Harari, A model for discovering "containment" relations, *Cognition* 183 (2019). https://doi.org/10.1016/j.cognition.2018.11.001.

[44] N. Dorfman, D. Harari, S. Ullman, Learning to perceive coherent objects, in *Proc. 35th Annu. Meet. Cogn. Sci. Soc.* (2013).

[45] M. Casasola, L.B. Cohen, E. Chiarello, Six-month-old infants' categorization of containment spatial relations, *Child Dev.* 74 (2003). https://doi.org/10.1111/1467-8624.00562.

[46] S. Ullman, Using neuroscience to develop artificial intelligence, *Science* 363 (2019). https://doi.org/10.1126/science.aau6595.

[47] E. Borenstein, S. Ullman, Class-specific, top-down segmentation, in *Lect. Notes Comput. Sci.* (2003). https://doi.org/10.1007/3-540-47967-8_8.

[48] E. Borenstein, E. Sharon, S. Ullman, Combining top-down and bottom-up segmentation, in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work* (2004). https://doi.org/10.1109/CVPR.2004.314.

[49] E. Borenstein, S. Ullman, Combined top-down/bottom-up segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 30 (2008). https://doi.org/10.1109/TPAMI.2007.70840.

[50] S. Ullman, Object recognition and segmentation by a fragment-based hierarchy, *Trends Cogn. Sci.* 11 (2007). https://doi.org/10.1016/j.tics.2006.11.009.

[51] T.D. Albreight, On the perception of probable things: Neural substrates of associative memory, imagery and perception. *Neuron* 74 (2012). https://doi.org/10.1016/j.neuron.2012.04.001.

[52] S. Ullman, L. Assif, A. Strugatski, B. Vatashsky, H. Levi, A. Netanyahu, A. Yaari, Image interpretation by iterative bottom-up top-down processing, *Proc. Natl. Acad. Sci. U.S.A.* 120 (2021). http://arxiv.org/abs/2105.05592.

[53] D.L.K. Yamins, H. Hong, J.J. DiCarlo C.F. Cadieu, Performance-optimized hierarchical models predict neural responses in higher visual cortex, *Proc. Natl. Acad. Sci. U.S.A.* 111 (2014) https://doi.org/10.1073/pnas.1403112111.

[54] M. Schrimpf, J Kubilius, M.J Lee, N.A.R Murty, R Ajemian, J.J DiCarlo, Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108 (2020). https://doi.org/10.1016/j.neuron.2020.07.040.