Deep Learning of Neuromuscular and Visuomotor Control of a Biomimetic Simulated Humanoid

Masaki Nakada, Tao Zhou, Honglin Chen, Arjun Lakshmipathy, and Demetri Terzopoulos, *Fellow, IEEE*

***Abstract*—We present a biomimetic framework for human neuromuscular and visuomotor control that promises to be of value to researchers developing humanoid robots. Our framework features a biomechanically simulated human musculoskeletal model, actuated by numerous skeletal muscles, with realistic eyes driven by extraocular and intraocular muscles, whose optic organs refract light, and whose retinas have many nonuniformly distributed photoreceptors. The humanoid's visuomotor control system comprises 24 trained deep neural networks (DNNs)—10 DNNs in its vision subsystem and 14 DNNs in its motor subsystem—plus an additional 4 trained shallow neural networks (SNNs) that control the irises and lenses of the eyes. Of the motor DNNs, a pair control the extraocular muscles, 6 per eye, responsible for eye movements, 2 control the 216 neck muscles of the cervicocephalic biomechanical complex, producing natural head movements, 2 control the 443 core muscles of the torso, and 2 control each limb; i.e., the 29 muscles of each arm and 39 muscles of each leg. Directly from the foveated retinal photoreceptor responses, a pair of foveation DNNs drive eye, head, and torso movements, while 8 limb vision DNNs extract the visual information needed to direct arm and leg actions. By synthesizing its own training data, our humanoid automatically learns efficient, online, active visuomotor control of its eyes, head, torso, and limbs in order to perform nontrivial tasks involving the foveation and visual pursuit of moving target objects coupled with visually-guided limb-reaching actions to intercept them. We also demonstrate that it can balance itself in an upright stance, take steps, and perform certain simulated sports activities.**

***Index Terms*—Biomimetics, Simulation and Animation, Deep Learning in Robotics and Automation, Visual Learning, Humanoid Robots**

I. Introduction

Anthropomorphic physical robots, in some sense the "Holy Grail" of robotics, have been pursued by several research groups [1], [2], [3], [4], [5], [6]. When musculoskeletal structures are attempted, they usually are highly simplified versions of their biological analogs, but researchers have shown that such biomimetic robotic systems can work. For example, Kurumaya *et al.* [7], [8] developed a humanoid robot with a musculoskeletal structure employing multi-filament muscles that have characteristics not unlike human muscles.

Visuomotor functionality in biological organisms refers to the acquisition and processing of visual input and the production of appropriate motor output responses to perform desired tasks. Visuomotor systems have also been explored for several decades in robotics [9], [10], [11], [12], [13]. Again, biomimetic approaches have been severely challenged by hardware constraints, but there have been some promising attempts in recent years. Bjorkman and Kragic [14] and Kragic *et al.* [15] introduced foveated visual processing using two binocular cameras with narrow and wide fields of view to approximate foveal and peripheral vision. Lesmana and Pai [16] and Lesmana *et al.* [17] employed a machine learning approach to generate robotic eye saccade, pursuit and VOR movements with oculomotor control systems.

In this paper, we present a strongly biomimetic *simulation* framework for human visuomotor control that is free of hardware constraints. Our framework is unique in that it features a biomechanically simulated human musculoskeletal model that currently includes no fewer than 823 anatomically accurate skeletal muscles, which actuate head, torso, arm, and leg motions, complemented by a pair of biomechanical eyes each with 2 intraocular muscles (iris sphincter and lens ciliary muscle) for visual accommodation plus 6 extraocular muscles responsible for realistic eye movements. Our simulated humanoid is a greatly enhanced version of a preliminary model [18] that had a non-functional torso (immobilized spine) and simplistic, kinematic eyes.

Our biomimetic visuomotor control system is unprecedented both in its use of a sophisticated biomechanical human model and in its use of modern machine learning methodologies. A modular set of neural networks controls the realistic musculoskeletal system and performs online visual processing for active, foveated perception and neuromuscular motor control. The neural networks are automatically trained from data synthesized by the human model itself.

Our muscle-actuated model represents a major advance in realism over the previous generation of simulated humanoids driven by rotational, "servomotor" joint actuators (e.g., [19]). Hence, our work should be of value to researchers developing sophisticated humanoid robot control systems, particularly since it demonstrates, through the synergistic combination of compatibly biomimetic motor and sensory neural controllers, what is possible with a simulated humanoid whose level of realism greatly exceeds current robotic hardware systems.

Furthermore, the benefit of a biomimetic simulated humanoid operating within in its physics-based virtual environment is that it can potentially provide a high-fidelity testbed for future anthropomimetic robots, offering a quick and inexpensive approach to developing and testing new biologically-inspired sensorimotor control theories, thereby accelerating progress.

Manuscript received September 10, 2019; revised December 12, 2019; accepted January 11, 2020.

This paper was recommended for publication by Editor Nancy Amato upon evaluation of the Associate Editor and Reviewers' comments. The work reported herein was presented to the RA community by D. Terzopoulos in an ICRA 2019 keynote address, a privilege for which he is grateful to Gregory Dudek and the other conference organizers.

The authors are with the Computer Science Department, University of California, Los Angeles, CA 90095, USA.

Digital Object Identifier (DOI): see top of this page.

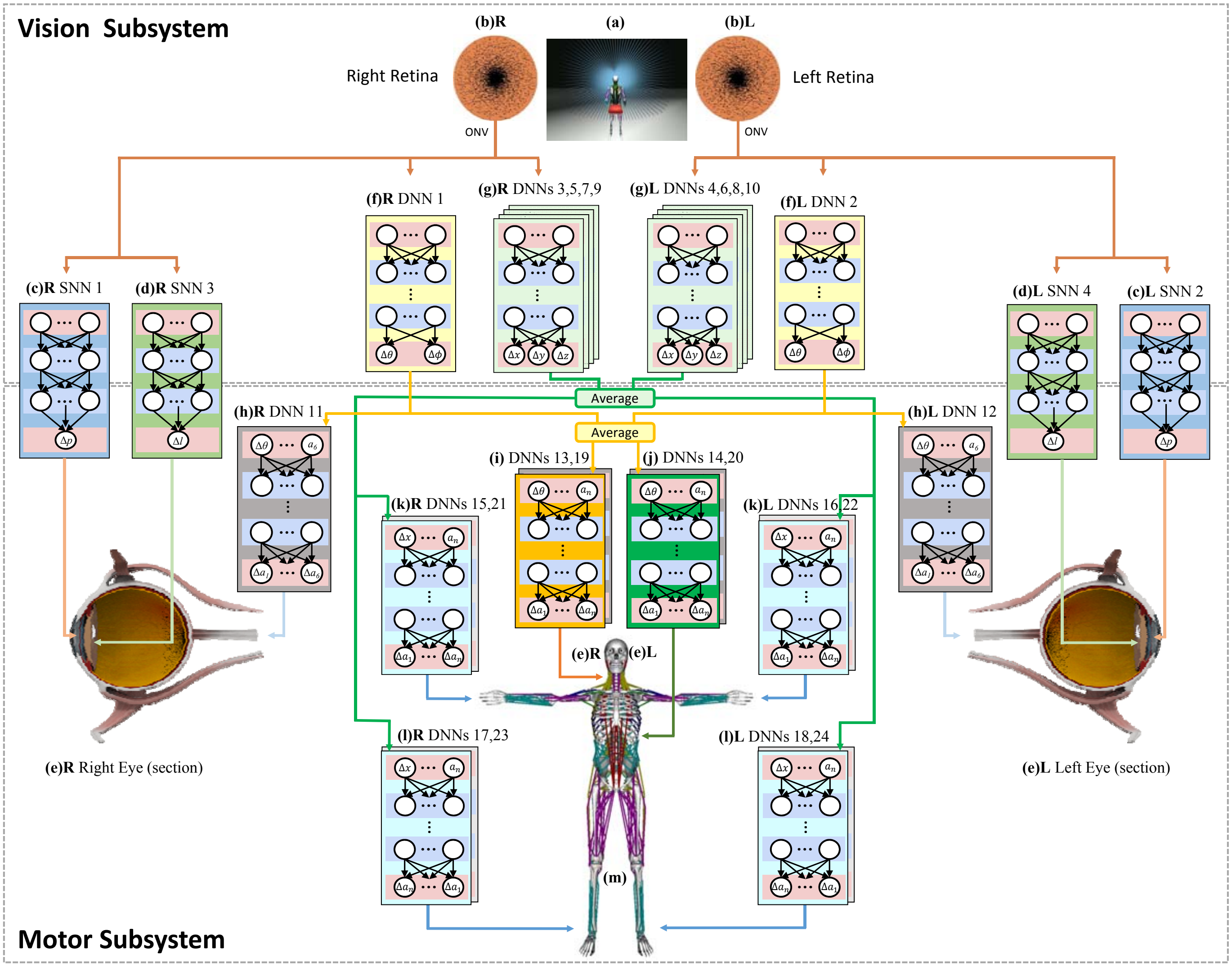


Fig. 1: Architecture of the visuomotor system. The controllers in the system include a total of 24 DNNs and 4 SNNs (Fig. 3).

Vision Subsystem (top): Each retinal photoreceptor casts rays through the eye (Fig. 2a,b) and into the virtual world (a), which compute the irradiance at that photoreceptor. (b) The arrangement of the 14,400 photoreceptors (black dots) on the right R and left L foveated retinas. Each eye outputs a 43,200-dimensional RGB Optic Nerve Vector (ONV). This feeds four trained visual accommodation SNNs (1–4); SNNs (c)R (1) and (d)R (3) control the muscles of the iris and lens of the right eye (e)R, and SNNs (c)L (2) and (d)L (4) do the same for the left eye (e)L. The ONV also feeds ten trained vision DNNs (1–10). (f) A pair of foveation DNNs (1,2) produce outputs that drive the movements of the eyes to foveate visual targets. (g) The eight limb vision DNNs (3–10) — (g)R (3,5,7,9) for the right eye and (g)R (4,6,8,10) for the left eye — output observed limb-to-target discrepancy estimates.

Motor Subsystem (bottom): Fourteen trained neuromuscular motor DNNs (11–24) comprise the motor subsystem, including eight voluntary motor DNNs (11–18) and six reflex motor DNNs (19–24). (h) The oculomotor DNNs (11,12), which are driven by the outputs of the foveation DNNs, output muscle activation signals that control the six extraocular muscles of each eye to produce eye movements. Driven by the averaged responses of the foveation DNNs, along with the current activations of the 216 neck muscles and 443 torso muscles, respectively, the cervicocephalic (i) voluntary motor DNN (13) and torso (j) voluntary motor DNN (14) each outputs muscle activation signals that contribute to actuating its associated neuromuscular complex. Driven by the bilaterally pairwise averaged responses of the limb vision DNNs, along with the current activations of the 29 muscles of each arm or 39 muscles of each leg, respectively, each of the four limb voluntary motor DNNs (k) (l) (15–18) outputs muscle activation signals that contribute to actuating its associated neuromuscular complex. Each of the six reflex motor DNNs (19–24) outputs muscle activation signals that contribute by stabilizing the muscle group of its associated musculoskeletal complex (Fig. 4).

II. Humanoid Model

As shown in Fig. 1, the visuomotor control system of our human model consists of a set of 24 automatically-trained, fully-connected Deep Neural Networks (DNNs) that operate continuously and synergistically, 14 of which are motor control DNNs in the motor subsystem, while the other 10 are vision DNNs in the vision subsystem. Additionally, 4 Shallow Neural Networks (SNNs) associated with the biomimetic eyes, suffice to control focal accommodation by activating the lens ciliary muscles as well as the iris sphincter muscles to regulate the amount of light that reaches the foveated retinas and their human-like, irregular photoreceptor distributions.

Directly from the retinal photoreceptor responses, a pair of foveation DNNs in the vision subsystem (upper Fig. 1) drive eye, head, and torso movements, while eight limb vision DNNs extract the perceptual information needed to drive the actions of the limbs. In the motor subsystem (lower Fig. 1), a pair of oculomotor DNNs control the 6 extraocular muscles per eye to perform eye movements. Additionally, two motor DNNs control the 216 neck muscles that actuate the cervicocephalic musculoskeletal complex to balance the head atop the flexible cervical column and produce controlled head movements. Two more control the 443 core muscles of the thoracic/lumbar musculoskeletal complex that support and actuate the torso. Finally, eight motor DNNs control the musculoskeletal complexes of the four limbs; in particular, the 29 muscles of each arm and the 39 muscles of each leg.

Thus, our simulated humanoid is capable of learning efficient, online visuomotor control of its eyes, head, torso, and four limbs to perform nontrivial motor tasks driven exclusively by its egocentric, active visual perception.

A. *Biomechanical Body Model*

Fig. 1m displays the musculoskeletal system of our anatomically accurate human model, which includes all of the relevant articular bones and muscles—193 bones connected by joints comprising 163 articular degrees of freedom, plus a total of 823 muscle actuators embedded in a finite element model of the musculotendinous soft tissues of the body.[1] We designed our biomechanical humanoid's musculoskeletal system by referring to one of the most comprehensive, albeit purely geometric, commercially available human models, the Ultimate Human Model.[2] Biomechanical parameters and muscle attachment points were determined using the geometric model as a reference.[3] Each skeletal muscle is modeled as a Hill-type uniaxial contractile actuator that applies forces to the bones at its points of insertion and attachment. The human model is numerically simulated as a force-driven articulated multi-body system.[4]

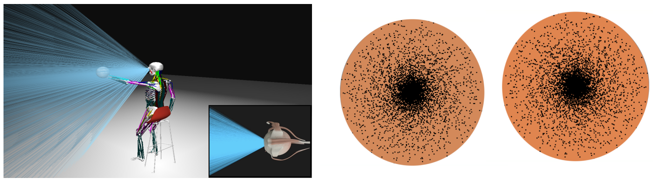

(a) (b) (c) Left retina (d) Right retina

Fig. 2: Via raytracing (blue rays), the biomechanical humanoid's eyes (b) visually sample the environment (a) to compute the irradiance at the locations of RGB photoreceptors (black dots) placed according to a foveated, noisy log-polar distribution on their retinas (c) (d).

Each muscle actuator is activated by an independent, time-varying, efferent activation signal $a(t)$. Given our human model, the overall challenge in neuromuscular motor control is to determine the activation signals for each of its 823 muscles necessary to carry out various motor tasks.

B. *Biomechanical Eye Model*

We next summarize our biomimetic eye model, the details of which are presented in [23].

To model the eyes, we have taken into consideration physiological data. As shown in Fig. 1e, we model the virtual eye as a sphere of radius 12mm with the typical 7.5 g mass of the human eyeball, which can be rotated with respect to its center around its vertical y axis by a horizontal angle of θ, around its horizontal x axis by a vertical angle of ϕ, and around the gaze z axis by a torsion angle ψ. The eyes look forward in their neutral positions when $\theta = \phi = \psi = 0°$.

As with the human eye, each virtual eyeball is actuated by 6 extraocular muscles, including the 4 rectus muscles that actuate much of the θ-ϕ movement, and the 2 oblique muscles that induce torsion.

Our biomimetic eye model includes a cornea that refracts light rays, an iris with a finite-aperture pupil capable of dilation and constriction to regulate the incoming light, and a model of the ocular lens that further refracts light rays and, through active lens deformation, can focus them onto the retinal surface. See [23] for additional details.

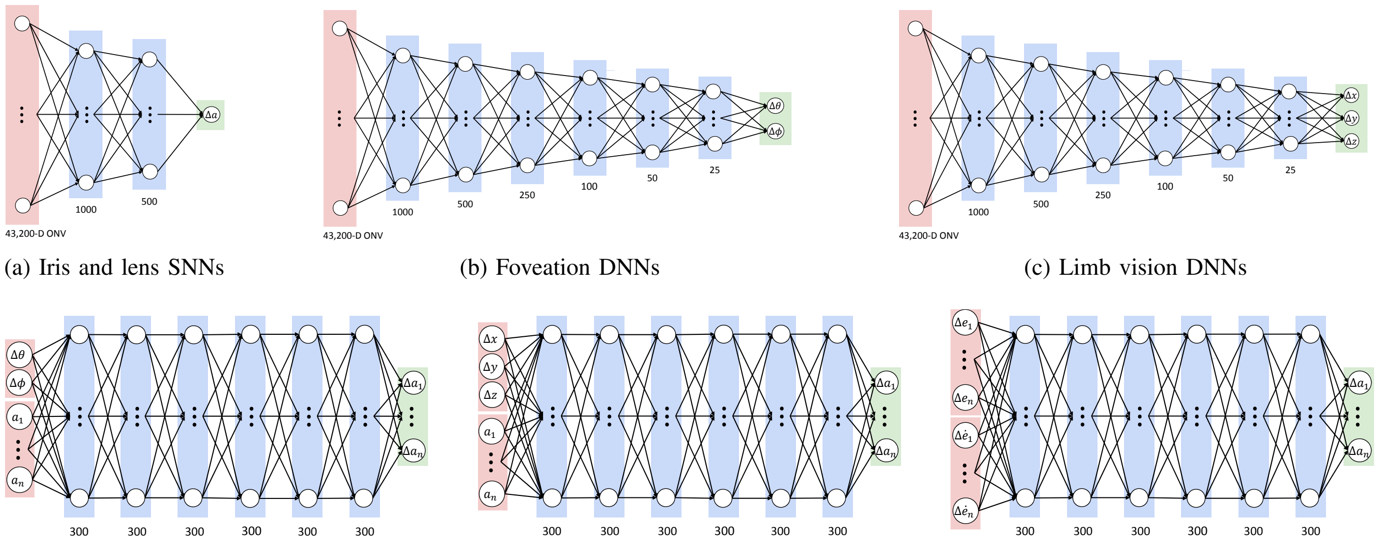
1) Model of the Retina: The retina is the hemispherical inner surface at the rear of the eyeball. Our virtual retina emulates how biological retinas sample scene radiance from the incidence of light on photoreceptors. The irradiance at the locations of the photoreceptors, is computed using the raytracing technique of computer graphics rendering [24]. Fig. 2a,b illustrates the retinal "imaging" process. Each photoreceptor gathers light from the environment through the finite-aperture pupil, as follows: Sample rays emanate from the position of the photoreceptor, are refracted at the lens and corneal surfaces, and cast into the 3D virtual world where they recursively intersect with the visible surfaces of virtual objects and sample the virtual

[1] The finite element soft-tissue simulation, which produces realistic flesh deformations, is unnecessary in the scope of the present paper and is suppressed; however, the skeletal mass distribution remains that of an adult male body.

[2] https://www.turbosquid.com/3d-models/3d-human-anatomy-ultimate/1093983

[3] We opted against using the oversimplified biomechanical human models available to the community, such as OpenSim [20], which is not a whole-body model, and Anybody [21], a whole-body model with only 458 muscles. Per the Library of Congress (http://id.loc.gov/authorities/subjects/sh85088687.html), the human body has 650 named skeletal muscles and as many as 840 including unnamed muscles.

[4] For the details, see [22] and Appendix I. Appendices I, II, and III are found in the supplemental document.



(a) Iris and lens SNNs (b) Foveation DNNs (c) Limb vision DNNs

(d) Ocular, cervicocephalic, and torso voluntary motor DNNs (e) Limb voluntary motor DNNs (f) Reflex motor DNNs

Fig. 3: Network architectures of the (a)–(c) vision SNNs and DNNs, and (d)–(f) motor DNNs.

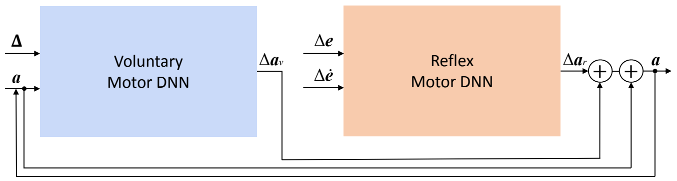


Fig. 4: Architecture of a neuromuscular motor controller. The voluntary motor DNN inputs the discrepancy $\boldsymbol{\Delta}$ and, recurrently, the muscle activations $\boldsymbol{a}$, and outputs muscle activation updates $\Delta\boldsymbol{a}_v$, which induce the desired actuation of the associated musculoskeletal complex. The reflex motor DNN inputs the discrepancies in muscle strains $\boldsymbol{e}$ and strain rates $\dot{\boldsymbol{e}}$, and outputs stabilizing activation updates $\Delta\boldsymbol{a}_r$. The output of the controller is $\boldsymbol{a}(t+\Delta t) = \boldsymbol{a}(t) + (\Delta\boldsymbol{a}_v(t) + \Delta\boldsymbol{a}_r(t))$.

light sources in accordance with the standard Phong local illumination model. The total irradiance at each photoreceptor is the weighted sum of the associated sample ray irradiances.

Unlike the uniform, Cartesian grid arrangement of most robot imaging sensors, visual sampling in the primate retina is known to be strongly space-variant with the density of trichromatic (cone) photoreceptors decreasing radially from the fovea toward the periphery [25]. The log-polar distribution is often used in space-variant image sampling models. To simulate biomimetic foveated perception, we position the photoreceptors on the hemispherical retina according to a noisy log-polar distribution. On each retina (Fig. 2c,d), we include 14,400 RGB photoreceptors.

2) Optic Nerve Vectors: The foveated retinal RGB "image" captured by each eye is output for further processing down the visual pathway, not as an array of pixels, but as a 14,400 × 3 = 43,200-dimensional vector of photoreceptor responses, which we call the Optic Nerve Vector (ONV). The raw visual information encoded in the ONV supplies the vision DNNs. These directly control eye movements and extract perceptual information that is passed on to the neuromuscular motor control DNNs in the motor subsystem, which drive head and torso movements as well as the reaching actions of the limbs.

III. Visuomotor System

Fig. 1 overviews the visuomotor system, detailing its vision and motor subsystems. The figure caption describes the information flow and functions of its 24 DNN controllers (numbered 1–24 in the figure) and 4 SNNs. Fig. 3 illustrates the architectures of the various networks, which will be explained in the ensuing sections.[5]

The vision subsystem includes 10 fully-connected, feedforward DNNs, numbered 1–10 in Fig. 1, which process the visual information in the 43,200-dimensional ONV. These DNNs are of two types. The first, described in Section III-A1, controls the eye movements, as well as the head and torso movements via the neck and torso neuromuscular motor controller. The second type, described in Section III-A2, produces arm-to-target 3D discrepancies $[\Delta x, \Delta y, \Delta z]$ that drive the limbs via the limb neuromuscular motor controllers.

The motor subsystem includes a pair of oculomotor controllers and six neuromuscular motor controllers associated with the cervicocephalic, torso, and four limb musculoskeletal

[5] As is detailed in Appendix III, we conducted experiments with various neural network architectures, activation functions, and other parameters to determine their suitability for our purposes. The 6-hidden-layer architecture common to all our DNNs has proven in our experiments to provide the best overall performance. All the networks illustrated in Fig. 3 employ rectified linear units (ReLU). Their initial weights are sampled from the zero-mean normal distribution with standard deviation $\sqrt{2/fan\_in}$, where fan_in is the number of input units in the weight tensor [26]. Network unit/weight counts: (a)–(c) ∼45K/∼44M; (d)–(e) ∼2K/∼400K–700K; (f) ∼2K–3K/∼400K–800K.

The DNN training data synthesis procedures are described in the ensuing sections and in Appendix II. To train the networks, we use the mean-squared-error loss function and the Adaptive Moment Estimation (Adam) [27] stochastic optimizer with learning rate $\eta = 10^{-6}$, step size $\alpha = 10^{-3}$, forgetting factors $\beta_1 = 0.9$ for gradients and $\beta_2 = 0.999$ for second moments of gradients. Overfitting is avoided using an early stopping condition: negligible improvement for 10 successive epochs.

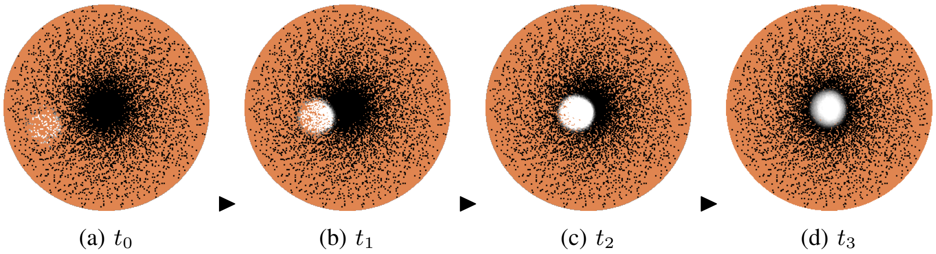


Fig. 5: Time sequence of photoreceptor responses in the left retina during a saccadic eye movement that foveates and tracks a moving white ball — (a) at time t_0 the ball enters the visual periphery, (b) at t_1 the eye movement is bringing the ball towards the foveal region, (c) at time t_2 the moving ball is centering within the foveal region, (d) at time t_3 the ball is foveated/tracked.

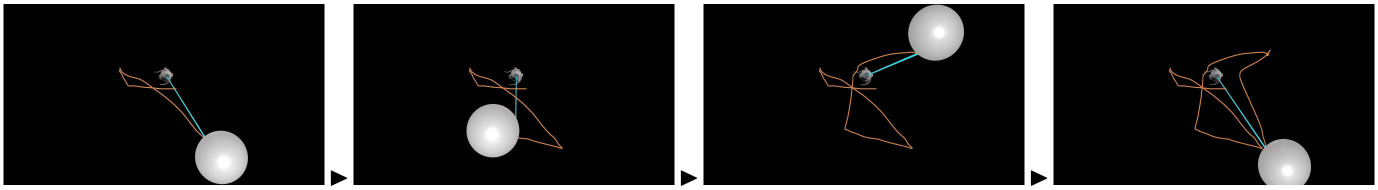


Fig. 6: Simulation of saccadic eye movement dynamics driven by the oculomotor DNNs. Eye movement (red trace) from one gaze direction to the next (blue rays) is triggered by sudden transitions of the visual target (white ball).

complexes. Fig. 4 illustrates the neuromuscular motor controller architecture, which is comprised of a voluntary motor DNN and a reflex motor DNN (c.f. [28]), both of which are fully-connected, feedforward DNNs that produce muscle activation adjustment signals, $\Delta\boldsymbol{a}$. The muscle activation $\boldsymbol{a}$ feedback loop makes the neuromuscular controllers recurrent networks. The voluntary motor DNNs, numbered 11–18 in Fig. 1, are discussed in Sections III-B1 to III-B4, and the reflex motor DNNs, numbered 19–24, are discussed in Section III-B5. The oculomotor controllers, DNNs 11–12, do not require stabilization from reflex motor DNNs.

A. Vision Subsystem DNNs

1) Foveation DNNs (1,2): The first role of the left and right foveation DNNs is to drive saccadic eye movements that alter the gaze directions in order to foveate visible objects of interest, thereby observing them with maximal visual acuity, as illustrated in Fig. 5 for a white ball in motion that enters the field of view stimulating a number of peripheral photoreceptors on the retina. The eye performs a rapid saccadic rotation to foveate the visual target, as shown in Fig. 6. Fine adjustments comparable to microsaccades are observed during fixation, which is followed by smooth pursuit eye movement as the foveated target moves.

The second role of the foveation DNNs is to control head and torso movements to facilitate fixation and visual tracking. This is accomplished simply by driving the cervicocephalic and torso neuromuscular motor controllers (DNNs 13,19,14,20 in Fig. 1i,j) with the averaged outputs of the foveation DNNs. The dynamic head and torso movements are thus coupled to eye movements induced by visual attention.

As Fig. 3b reveals, the input layer of a foveation DNN comprises 43,200 units, to accommodate the dimensionality of the ONV, the output layer has 2 units, $\boldsymbol{\Delta} = [\Delta\theta, \Delta\phi]^T$, and there are six hidden layers. Our humanoid trains its foveation networks, as follows: A white ball is presented within the visual field. By raytracing the 3D scene from the perspective of the eye, the photoreceptors on its retina are stimulated and their RGB components comprise the ONV input to the network. The desired output of a foveation DNN is the angular differences $\boldsymbol{\Delta}$ between the actual gaze directions of the eyes and the known gaze directions that would foveate the ball. Repeatedly positioning the visual target at random locations in the visual field synthesizes a large training dataset of 1M input-output pairs offline. The backpropagation-trained DNN serves as an online foveation controller.

2) Limb Vision DNNs (3–10): The role of the left and right limb (arm and leg) vision DNNs is to estimate the separation in 3D space between the position of the end effector (hand or foot) and the position of a visual target, thus driving the associated limb motor DNN to extend the limb in order to touch the target.

The architecture of the limb vision DNNs (Fig. 3c) is identical to the foveation DNNs except for the size of the output layer, which has 3 units $\boldsymbol{\Delta} = [\Delta x, \Delta y, \Delta z]^T$, to encode the estimated discrepancy between the 3D positions of the end effector and the visual target.

Our humanoid trains its four limb vision DNNs, as follows: A ball is presented in the visual field and the trained foveation DNNs foveate the visual target. Then, a limb (arm or leg) is extended towards the ball. The retinal photoreceptors in the eyes are stimulated and the visual stimuli are presented as the RGB components of the ONVs. Given its ONV input, the desired output of a limb vision DNN is the 3D discrepancy $\boldsymbol{\Delta}$ between the known 3D positions of the end effector and visual target. Repeatedly placing the ball at random positions in the visual field and articulating the limb to reach for it in space, a large training dataset of 1M input-output pairs is synthesized offline. The backpropagation-trained DNN serves as an online limb vision controller.

B. Motor Subsystem DNNs

1) Oculomotor DNNs (11–12): For the oculomotor DNN controllers (Fig. 3d), the input layer consists of 8 units, 2 units for the angular discrepancies $\boldsymbol{\Delta} = [\Delta\theta,\ \Delta\phi]^T$ to the visual target and $n = 6$ units for the activations $\boldsymbol{a} = [a_1, \dots, a_n]^T$, of the six extraocular muscles. The output layer consists of 6 units providing the muscle activation adjustments $\Delta\boldsymbol{a} = [\Delta a_1, \dots, \Delta a_n]^T$. Each of the six hidden layers contains 300 units.

To train the DNNs, our biomechanical humanoid synthesizes training data as follows: Specifying a target orientation for the eye yields angular discrepancies between it and the current eye orientation. With the angular discrepancies $\boldsymbol{\Delta}$ and current extraocular muscle activations $\boldsymbol{a}$ as input, the biomechanical eye model computes inverse dynamics with minimal effort optimization of the muscles. This determines muscle activation adjustments $\Delta\boldsymbol{a}$ that incrementally reduce the angular discrepancies and serve as the desired output of the oculomotor DNN. Repeatedly specifying random target eye orientations, a large training dataset of 1M input-output pairs was synthesized offline. The backpropagation-trained DNN serves as online oculomotor controllers.

2) Cervicocephalic Voluntary Motor DNN (13): Similarly, the input layer of the cervicocephalic voluntary motor DNN (Fig. 3d) consists of 218 units, 2 units for the head pose target discrepancy angles $\boldsymbol{\Delta} = [\Delta\theta, \Delta\phi]^T$ and $n = 216$ units for the activations $\boldsymbol{a}$, of the 216 muscles of the cervicocephalic musculoskeletal complex. The output layer consists of 216 units providing the muscle activation adjustments $\Delta\boldsymbol{a}_v$. Each of the six hidden layers contains 300 units.

To train the DNN, our biomechanical humanoid synthesizes training data as follows: Specifying a target orientation for the head yields angular discrepancies between it and the current head orientation. With the angular discrepancies $\boldsymbol{\Delta}$ and current neck muscle activations $\boldsymbol{a}$ as input, our humanoid computes inverse kinematics (IK), inverse dynamics (ID), and muscle optimization (MO).[6] This determines muscle activation adjustments $\Delta\boldsymbol{a}_v$ that incrementally reduce the angular discrepancies and serve as the desired output of the cervicocephalic DNN. Repeatedly specifying random angular discrepancies, a training set of 1M input-output training pairs was synthesized offline. The backpropagation-trained DNN serves in the online cervicocephalic neuromuscular motor controller.

3) Core Voluntary Motor DNN (14): The core neuromuscular controller voluntary motor DNN (Fig. 3d) is architected identically to the cervicocephalic DNN, except for the size of the input and output layers. The input layer consists of 446 units that comprise 3 units for the T1 vertebra target angular discrepancies $\boldsymbol{\Delta} = [\Delta\alpha, \Delta\beta, \Delta\gamma]^T$ and $n = 443$ units for the activations $\boldsymbol{a}$ of the muscles of the torso musculoskeletal complex. The output layer consists of 443 units providing the muscle activation adjustments $\Delta\boldsymbol{a}_v$. The training proceeds similarly to that for the cervicocephalic voluntary motor DNN, but see Section III-B6.

4) Limb Voluntary Motor DNNs (15–18): The input layer of the limb voluntary motor DNNs (Fig. 3e) consists of 3 units that specify $\boldsymbol{\Delta} = [\Delta x,\ \Delta y,\ \Delta z]^T$, the estimated discrepancy between the 3D positions of the end effector and a specified target position, as well as the current activations $\boldsymbol{a}$ of the $n = 29$ arm muscles or those of the $n = 39$ leg muscles. The output layer consists of equal numbers of units that encode the muscle activation adjustments $\Delta\boldsymbol{a}_v$.

To train the DNNs, our biomechanical humanoid again synthesizes training data: Specifying a target position, determines the discrepancy between it and the end effector (hand or foot). Given the discrepancy $\boldsymbol{\Delta}$ and current muscle activations $\boldsymbol{a}$ as the input, the desired output of the network is the muscle activation adjustments $\Delta\boldsymbol{a}_v$, which are again computed through IK followed by ID and MO of the limb muscles. Repeatedly specifying random target positions, a large training dataset of 1M input-output training pairs was synthesized offline. The backpropagation-trained DNNs serve in the limb online neuromuscular motor controllers.

5) Reflex Motor DNNs (19–24): Fig. 3f illustrates the architecture of the 6 reflex motor DNNs. They are identical, except for the sizes of their input and output layers, which are determined by the number n of muscles in the associated musculoskeletal complex. The input layer consists of $2n$ units, which represent the change in muscle strains $\Delta\boldsymbol{e} = [\Delta e_1, \dots, \Delta e_n]^T$ and strain rates $\Delta\dot{\boldsymbol{e}} = [\Delta\dot{e}_1, \dots, \Delta\dot{e}_n]^T$. Like the voluntary motor DNNs, the networks have six hidden layers with 300 units each. The output layer consists of n units providing muscle activation adjustments $\Delta\boldsymbol{a}_r$, which then additively modify the muscle activations.

Training data for the reflex motor DNNs are computed offline simultaneously with the associated voluntary motor DNN training data synthesis. The input discrepancies in muscle strains $\Delta\boldsymbol{e}$ and strain rates $\Delta\dot{\boldsymbol{e}}$, are computed after the IK phase, and the desired $\Delta\boldsymbol{a}_r$ output is computed as a proportional-derivative (PD) controller. Because this computation is on a per-muscle basis, the ID and MO phases are unnecessary. The backpropagation-trained reflex motor DNNs serve in the online neuromuscular motor controllers.

6) Core Training: Going beyond the independent training of the five decoupled extremities in our preliminary biomechanical musculoskeletal model [18], which lacked a functional torso, the six musculoskeletal complexes (torso and extremities) must be regarded a unified whole in order to properly train the core neuromuscular motor controller of the torso musculoskeletal complex. This is because the articulated biomechanical skeletal structure remains connected when in motion and each of the musculoskeletal complexes of the extremities include multiple significant muscles that attach to major bones in the torso. Hence, during torso training data synthesis, random muscle-actuated motions of the extremities are induced, including head turning, arm reaching, and leg squatting, such that forces from the extremities propagate to the torso. Furthermore, if the center of pressure approaches the boundary of the support polygon determined by the feet, the biomechanical model is reset to an

[6] More specifically, given $\boldsymbol{\Delta}$, IK yields the desired change of each of the vertebra joint angles using a quadratic programming solver, and the required joint accelerations are computed to obtain the desired angular modification within a specified time. Then, ID determines the desired joint torques to achieve the desired joint accelerations. Finally, the MO computation provides muscle activations that can produce the desired joint torques.

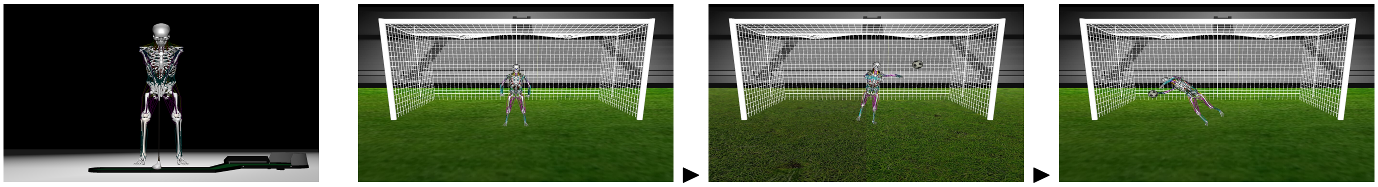

(a) Golf putting
(b) Balanced, ready stance
(c) Springing left
(d) Leaping right

Fig. 7: Our autonomous humanoid engaged in sports activities. (a) A frame from a simulated golf putting scenario. Holding the club in a "preacher" grip, the golfer steps forward and putts the ball into the hole. (b)–(d) A sequence of frames from a simulated soccer goaltending scenario. The goaltender observes the ball's trajectory, reaching out with its arms and leaping toward the ball as necessary in order to deflect it from the goal.

upright posture and data synthesis is restarted, thus training the core controller to maintain a balanced bipedal stance.[7]

IV. Simulation Results

Fig. 7 shows our autonomous humanoid with its visuomotor control system engaged in two sports activities—golf and soccer.[8] These were chosen to confirm that our humanoid can be dynamically controlled by its properly trained visuomotor system, while maintaining a balanced bipedal stance and demonstrating voluntary movement throughout its entire body, which was impossible for our earlier, constrained model [18].

Afforded binocular vision by virtue of its functional eyes as well as motor control to support the mass of its body and balance in gravity, our virtual golfer, shown in Fig. 7a, steps into position holding the putter with a "preacher" grip and putts the ball into the hole. As shown in Fig. 7b–d, balancing its body in an upright ready stance, our virtual soccer goaltender successfully observes a moving visual target, the incoming (massless) ball. Under neuromuscular oculomotor control, its eyes persistently track incoming soccer balls by making rapid, saccadic foveation and smooth pursuit eye movements, which in turn drive more sluggish (due to the greater masses involved) cervicocephalic and torso movements under neuromuscular musculoskeletal motor control. The goaltender's eyes track the moving ball while its head and body are in motion, which demonstrates Vestibulo-Ocular Reflex (VOR). Thus, by virtue of all its trained neuromuscular motor controllers operating in concert, our virtual goaltender controls its arms, and legs to reach out and even leap at approaching balls to deflect them from the goal.

Appendix A provides additional details regarding the control policies for the foregoing demonstrations.

V. Conclusions

We have presented a sophisticated simulation framework for exploring anthropomimetic visuomotor control. Our framework is unique in that it features an anatomically accurate, biomechanically simulated humanoid whose realistic skeleton is actuated by a full complement of contractile skeletal muscles. Furthermore, our model includes a pair of human-like eyes, each with a foveated retina accommodating numerous photoreceptors, as well as a functional cornea, iris sphincter, and deformable lens, which can adapt to illumination and achieve focus. Note that our model is a more complete version of the one presented in [18], which lacked a functional torso (immobilized lumbar and thoracic spine and pelvis) and utilized a simplistic eye model (kinematic, non-biomechanical eye movements and pinhole aperture).

Our elaborate anthropomimetic visuomotor system comprises a sizable collection of deep neuromuscular motor controllers and vision controllers. We have successfully demonstrated its robust performance in task scenarios that simultaneously involve eye movement control for saccadic foveation and smooth pursuit of visual targets in conjunction with appropriate dynamic head and torso motion control, plus visually-guided dynamic limb control producing natural arm and leg extension actions that enable the humanoid to intercept moving target objects and perform other sports actions. Our model is also capable of performing balanced stepping maneuvers, but continuous bipedal locomotion is best actuated and controlled by incorporating low-level Central Pattern Generator (CPG) neural circuits to produce rhythmic muscle activations [29].

In particular, our humanoid's biomimetic eyes and visuomotor system make it possible to autonomously synthesize natural head-eye behavior for robots. This is important for two reasons. First, because our muscle-actuated biomechanical eye models produce appropriate saccades and smooth pursuit movements, they automatically gaze at and track moving objects of interest, which provides humanoid robots active perception of the environment and goal-directed visual information suitable for taking appropriate actions. Second, this meets the expectations of people who interact with a robot, as eye and head behaviors are perhaps those most critical to human sensitivities. Among its many potential uses, our model promises to be valuable in human visual attention research, a topic that we wish to explore in future work. For this, as well as for other types of visual processing, such as binocular stereopsis, we will want

[7] The entire biomechanical body must be simulated during offline training data synthesis for the core neuromuscular motor controller; thus, on an Intel Xeon E5-1650 v4 3.60 MHz CPU, 8.7 s on average are required to compute each input-output training pair for its voluntary motor DNN and an additional 0.3 s for its reflex motor DNN. By comparison, the neuromuscular motor controllers of the extremities are trained independently; thus, for the cervicocephalic controller, the respective times are 17 ms and 2 ms, and for all four limb controllers, they are 4 ms and 1 ms. (Note that for the oculomotor controller, offline training data synthesis requires approximately 0.1 s per input-output pair on an Intel Core i7-6700 3.4 GHz CPU.) Online, the trained neuromuscular motor controllers compute muscle activations rapidly—the core controller in 2.8 ms, the cervicocephalic controller in 2.2 ms, and the limb controllers in 0.7 ms on the Xeon CPU.

[8] We refer the reader to our demonstration video.

to increase the number of photoreceptors, experiment with other nonuniform photoreceptor distributions, and automatically construct 2D retinotopic maps from the 1D ONV inputs.

Our visuomotor control system should, in principle, be transplantable into an anthropomorphic physical robotic system, were a compatible one available. However, building one would involve continuing research across multiple fields including materials science to create more highly biomimetic muscle actuators [7] and sensing technologies to create eye-like, foveated imaging devices [14], [15]. Although these are ambitious goals, our work reaffirms that they have great potential.

Appendix A
Control Policies

a) Golfing: We set "ideal" lower-body postures for three states (standing, stepping, and putting). Transitional movements between the states are generated naturally by the trained neuromuscular motor controllers. The control objective for the torso is to keep the body upright (with some natural forward-leaning angles) and balanced, as the center of pressure is monitored during the training of the core controller. The objective for the arms is to grip the club in the initial putting posture, and then swing back to a desired position. The objective for the eyes is to foveate the golf ball, and the cervicocephalic controller induces the head to follow the eye movement.

b) Soccer: The policies for the eye, cervicocephalic, and torso neuromuscular motor controllers are the same as for golfing. The objective for both arms is to reach a incoming ball observed by the eyes. Initially, the knees are slightly bent in the natural preparatory stance of a goaltender ready for fast reaction. When the eyes detect a ball approaching on one side of the body, the opposite-side leg neuromuscular motor controller triggers a rapid leg extension. This, in conjunction with the momentum of arm extensions generates leaping movements toward the incoming ball.

Learning higher-level control policies for performing optimal sports movements, possibly through reinforcement learning, is a good topic for future work.

References

[1] I. Mizuuchi, Y. Nakanishi, Y. Sodeyama, Y. Namiki, T. Nishino, N. Muramatsu, J. Urata, K. Hongo, T. Yoshikai, and M. Inaba, "An advanced musculoskeletal humanoid Kojiro," in *Proc. IEEE-RAS International Conference on Humanoid Robots*, 2007, pp. 294–299.

[2] H. G. Marques, M. Jäntsch, S. Wittmeier, O. Holland, C. Alessandro, A. Diamond, M. Lungarella, and R. Knight, "ECCE1: The first of a series of anthropomimetic musculoskeletal upper torsos," in *Proc. IEEE-RAS International Conference on Humanoid Robots*, 2010, pp. 391–396.

[3] K. Ogawa, K. Narioka, and K. Hosoda, "Development of whole-body humanoid "Pneumat-BS" with pneumatic musculoskeletal system," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 4838–4843.

[4] R. Niiyama, S. Nishikawa, and Y. Kuniyoshi, "Biomechanical approach to open-loop bipedal running with a musculoskeletal athlete robot," *Advanced Robotics*, vol. 26, no. 3-4, pp. 383–398, 2012.

[5] Y. Nakanishi, S. Ohta, T. Shirai, Y. Asano, T. Kozuki, Y. Kakehashi, H. Mizoguchi, T. Kurotobi, Y. Motegi, K. Sasabuchi, *et al.*, "Design approach of biologically-inspired musculoskeletal humanoids," *International Journal of Advanced Robotic Systems*, vol. 10, no. 4, pp. 216:1–13, 2013.

[6] Y. Asano, K. Okada, and M. Inaba, "Design principles of a human mimetic humanoid: Humanoid platform to study human intelligence and internal body system," *Science Robotics*, vol. 2, no. 13, pp. eaaq0899:1–11, 2017.

[7] S. Kurumaya, K. Suzumori, H. Nabae, and S. Wakimoto, "Musculoskeletal lower-limb robot driven by multifilament muscles," *Robomech Journal*, vol. 3, no. 18, pp. 1–15, 2016.

[8] S. Kurumaya, H. Nabae, G. Endo, and K. Suzumori, "Design of thin McKibben muscle and multifilament structure," *Sensors and Actuators A: Physical*, vol. 261, pp. 66–74, 2017.

[9] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 3, pp. 313–326, 1992.

[10] W. J. Wilson, C. W. Hulls, and G. S. Bell, "Relative end-effector control using Cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 684–696, 1996.

[11] M. Jagersand, O. Fuentes, and R. Nelson, "Experimental evaluation of uncalibrated visual servoing for precision manipulation," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 4, 1997, pp. 2874–2880.

[12] P. Hebert, N. Hudson, J. Ma, T. Howard, T. Fuchs, M. Bajracharya, and J. Burdick, "Combined shape, appearance and silhouette for simultaneous manipulator and object tracking," in *Proc. IEEE International Conference on Robotics and Automation*, 2012, pp. 2405–2412.

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[14] M. Bjorkman and D. Kragic, "Combination of foveal and peripheral vision for object recognition and pose estimation," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 5, 2004, pp. 5135–5140.

[15] D. Kragic, M. Björkman, H. I. Christensen, and J.-O. Eklundh, "Vision for robotic object manipulation in domestic settings," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 85–100, 2005.

[16] M. Lesmana and D. Pai, "A biologically inspired controller for fast eye movements," in *Proc. IEEE International Conference on Robotics and Automation*, 2011, pp. 3670–3675.

[17] M. Lesmana, A. Landgren, P.-E. Forssén, and D. Pai, "Active gaze stabilization," in *Proc. Indian Conference on Computer Vision, Graphics and Image Processing*, 2014, pp. 81:1–8.

[18] M. Nakada, T. Zhou, H. Chen, T. Weiss, and D. Terzopoulos, "Deep learning of biomimetic sensorimotor control for biomechanical human animation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 56:1–14, 2018, (Proc. *ACM SIGGRAPH 2018*, Vancouver, Canada, August 2018).

[19] P. Faloutsos, M. Van De Panne, and D. Terzopoulos, "Autonomous reactive control for simulated humanoids," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 1, 2003, pp. 917–924.

[20] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "OpenSim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[21] J. B. Langholz, G. Westman, and M. Karlsteen, "Musculoskeletal modelling in sports-evaluation of different software tools with focus on swimming," *Procedia Engineering*, vol. 147, pp. 281–287, 2016.

[22] S.-H. Lee, E. Sifakis, and D. Terzopoulos, "Comprehensive biomechanical modeling and simulation of the upper body," *ACM Transactions on Graphics*, vol. 28, no. 4, pp. 99:1–17, Aug. 2009.

[23] M. Nakada, A. Lakshmipathy, H. Chen, N. Ling, T. Zhou, and D. Terzopoulos, "Biomimetic eye modeling and deep neuromuscular oculomotor control," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 221:1–14, 2019, (Proc. *ACM SIGGRAPH Asia 19 Conference*, Brisbane, Australia, November 2019).

[24] P. Shirley and R. K. Morley, *Realistic Ray Tracing*, 2nd ed. Natick, MA, USA: A. K. Peters, Ltd., 2003.

[25] E. L. Schwartz, "Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] S.-H. Lee and D. Terzopoulos, "Heads up! Biomechanical modeling and neuromuscular control of the neck," *ACM Trans. Graphics*, vol. 23, no. 212, pp. 1188–1198, 2006.

[29] W. Si, S.-H. Lee, E. Sifakis, and D. Terzopoulos, "Realistic biomechanical simulation and control of human swimming," *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 10:1–15, Nov. 2014.