Biomimetic Perception Learning for Human Sensorimotor Control

Masaki Nakada, Honglin Chen, and Demetri Terzopoulos

University of California, Los Angeles

Abstract. We present a simulation framework for biomimetic human perception and sensorimotor control. It features a biomechanically simulated, musculoskeletal human model actuated by numerous skeletal muscles, with two human-like eyes whose retinas have spatially nonuniform distributions of photoreceptors. Our prototype sensorimotor system for this model incorporates a set of 20 automatically-trained, deep neural networks (DNNs), half of which are neuromuscular DNN controllers comprising its motor subsystem, while the other half are devoted to visual perception. Within the sensory subsystem, which continuously operates on the retinal photoreceptor outputs, 2 DNNs drive eye and head movements, while 8 DNNs extract the sensory information needed to control the arms and legs. Exclusively by means of its egocentric, active visual perception, our biomechanical virtual human learns efficient, online visuomotor control of its eyes, head, and four limbs to perform tasks involving the foveation and visual pursuit of target objects coupled with visually-guided reaching actions to intercept the moving targets.

Keywords: Sensorimotor control, Active vision, Foveated perception, Biomimetic vision, Deep learning.

1 Introduction

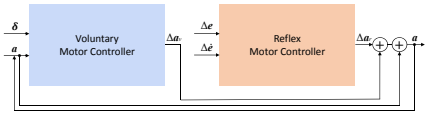
Biological vision has inspired computational approaches that mimic the functionality of neural mechanisms. Recent breakthroughs in machine learning with artificial (convolutional) neural networks have proven to be effective in computer vision; however, the application of Deep Neural Networks (DNNs) to sensorimotor systems has received virtually no attention in the computer vision field.

Sensorimotor functionality in biological organisms refers to the process of continually acquiring and interpreting sensory information necessary to produce appropriate motor responses in order to perform actions that achieve desired goals. We have recently introduced a simulation framework for investigating biomimetic human perception and sensorimotor control [5, 4]. Our framework is unique in that it features a biomechanically simulated, human musculoskeletal model, which currently includes 823 skeletal muscle actuators. Our virtual human perceives its environment using two eyes whose foveated retinas contain photoreceptors arranged in spatially nonuniform distributions.

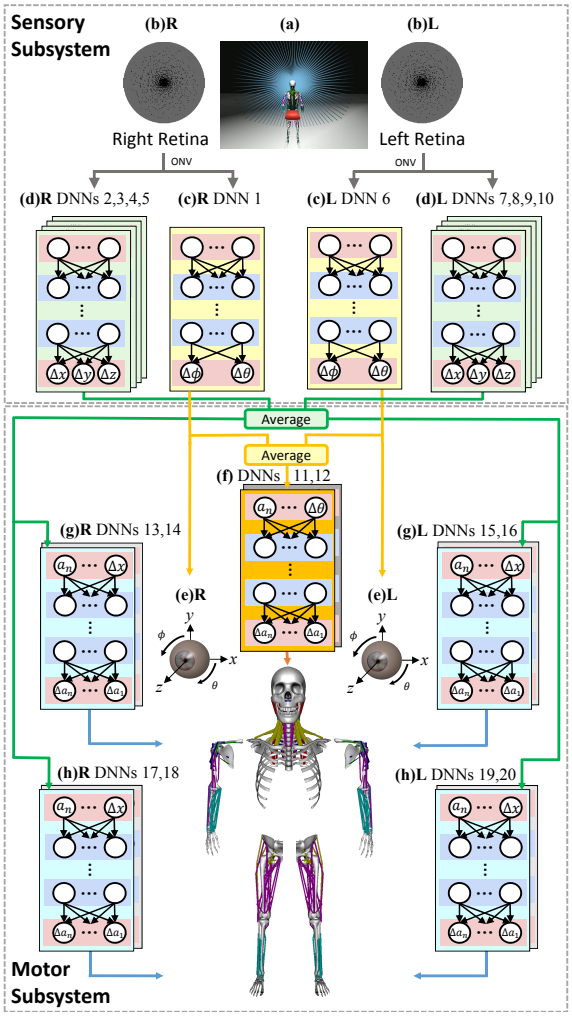
As illustrated in Fig. 1, we have developed a prototype visuomotor control system for our biomechanical human musculoskeletal model that incorporates

• Sensory Subsystem (top): (a) Each retinal photoreceptor casts a ray into the virtual world to compute the irradiance captured by the photoreceptor. (b) The arrangement of the photoreceptors (black dots) on the left (b)L and right (b)R foveated retinas. Each retina outputs an Optic Nerve Vector (ONV). There are 10 vision DNNs. The two (yellow) foveation DNNs (c) (1,6) input the ONV and produce eye movements to foveate visual targets. (d) The other eight (green) vision DNNs—(d)L (7,8,9,10) for the left eye (e)L and (d)R (2,3,4,5) for the right eye (e)R—also input the ONV and output limb-to-target visual discrepancy estimates.

• Motor Subsystem (bottom): There are 10 motor DNNs. The (orange) cervicocephalic neuromuscular motor controller (f) (DNNs 11,12) inputs the average of the foveation DNN responses and outputs activations to the neck muscle group. The four (blue) limb neuromuscular motor controllers (g),(h) (DNNs 13–20) of the limb musculoskeletal complexes input the average of the left (d)L and right (d)R limb vision DNN responses and output activations to the respective arm and leg muscle groups.



• Neuromuscular motor controller (above): The voluntary motor controller inputs a target discrepancy $\boldsymbol{\delta}$ and, recurrently, the muscle activations $\boldsymbol{a}$, and outputs adjustments $\Delta\boldsymbol{a}_v$ to the muscle activations that induce the desired actuation of the associated musculoskeletal complex. The reflex controller inputs the changes in muscle strains $\boldsymbol{e}$ and strain rates $\dot{\boldsymbol{e}}$ and produces muscle-stabilizing activation adjustments $\Delta\boldsymbol{a}_r$. The output of the neuromuscular motor controller is $\boldsymbol{a}(t+\Delta t) = \boldsymbol{a}(t) + (\Delta\boldsymbol{a}_v(t) + \Delta\boldsymbol{a}_r(t))$.



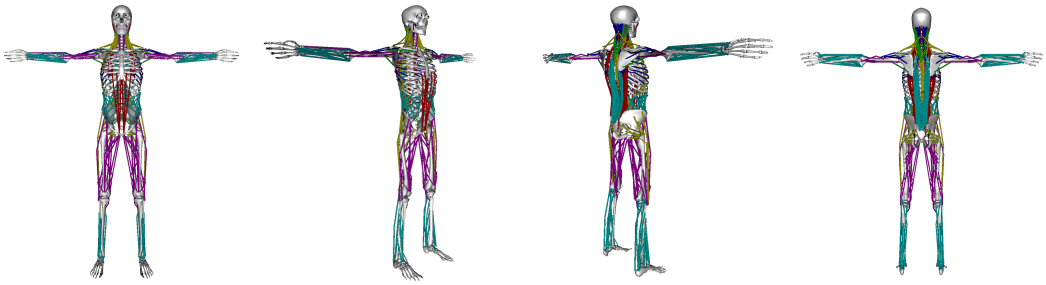


Fig. 1: The sensorimotor system architecture, whose controllers include a total of 20 DNNs, numbered 1–20, and the neuromuscular motor controller architecture. The complete biomechanical model (bottom), showing the skeletal system with its 103 bones and the 823 Hill-type muscle actuators.

a set of 20 automatically-trained, fully-connected DNNs, half of which are employed in neuromuscular motor controllers comprising its motor subsystem, while the other half are devoted to visual perception. Within its sensory subsystem (top of Fig. 1), which continuously operates on the retinal photoreceptor outputs, 2 vision DNNs drive eye and head movements, while 8 vision DNNs extract the sensory information needed to control the arms and legs. Thus, driven exclusively by means of egocentric, active visual perception, our biomechanical virtual human is capable of learning efficient and effective, online control of its eyes, head, and four limbs to perform visuomotor tasks. In particular, it demonstrates voluntary foveation and visual pursuit of target objects coupled with visually-guided reaching actions to intercept the moving targets.

Our work was inspired in part by the impressive visuomotor "EyeCatch" model of Yeo et al. [10]. However, EyeCatch is a non-learning-based visuomotor system embodied in a simple, non-physics-based humanoid character. By contrast, we have demonstrated dramatically more complex sensorimotor control realized using a set of trained DNNs in a comprehensive, anatomically-accurate, muscle-actuated biomechanical human model. Furthermore, the EyeCatch model and the more primitive visuomotor model of Lee and Terzopoulos [3] made direct use of the 3D spatial positions of virtual visual targets, without biologically-inspired visual processing. Instead, we have built upon the pioneering "Animat Vision" work on foveated, active computer vision for animated characters [9], which demonstrated vision-guided bipedal locomotion and navigation albeit in purely kinematic human characters [6]. In particular, we offer a substantially more biomimetic active vision model based on the foveated pattern of cone photoreceptor placement in biological retinas [7]. Given their fundamentally nonuniform distribution of photoreceptors, the retinas in our eye models capture the light intensity in the scene using raytracing, which better emulates how a biological retina samples scene radiance from the incidence of light on its photoreceptors.

Our visuomotor control system is unprecedented both in its use of a sophisticated biomechanical human model, as well as in its use of modern machine learning methodologies to control a realistic musculoskeletal system and perform online visual processing for active, foveated perception, all through a modular set of DNNs that are automatically trained from data synthesized by the human model itself.

The remainder of this paper is organized as follows: In Section 2, we overview our biomechanical human musculoskeletal model and its neuromuscular motor control subsystem (additional details are found in [5]). In Section 3 we describe its ocular and retinal models and in Section 4 we describe its sensory subsystem (additional details are found in [4]). Section 5 presents one of our experimental results. Section 6 summarizes our contributions and plans for future work.

2 Biomechanical Model and Motor Subsystem

The musculoskeletal system of our anatomically accurate biomechanical human model is shown at the bottom of Fig. 1. It includes all of the relevant articu-

lar bones and muscles—103 bones connected by joints comprising 163 articular degrees of freedom, plus a total of 823 muscle actuators. 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 (see [2] for the details).

Each muscle actuator is activated by an independent, time-varying, efferent activation signal $a(t)$. The overall challenge in the neuromuscular motor control of our human model is to determine the activation signals for each of its 823 muscles necessary to carry out various motor tasks. For the purposes of the present paper, we mitigate complexity by placing our virtual human in a seated position, immobilizing the pelvis as well as the lumbar and thoracic spinal column vertebra and other bones of the torso, leaving free to articulate only the cervicocephalic, two arm, and two leg musculoskeletal complexes (displayed within the lower box in Fig. 1).

The cervicocephalic musculoskeletal complex is rooted at the thoracic vertebra (T1), with its seven vertabrae, C7 through C1 (atlas), progressing up the cervical spine to the skull, which is an end-effector of substantial mass. A total of 216 short, intermediate, and long Hill-type uniaxial muscle actuators arranged in deep, intermediate, and superficial layers, respectively, actuate the seven 3-degree-of-freedom joints of the cervicocephalic musculoskeletal complex.

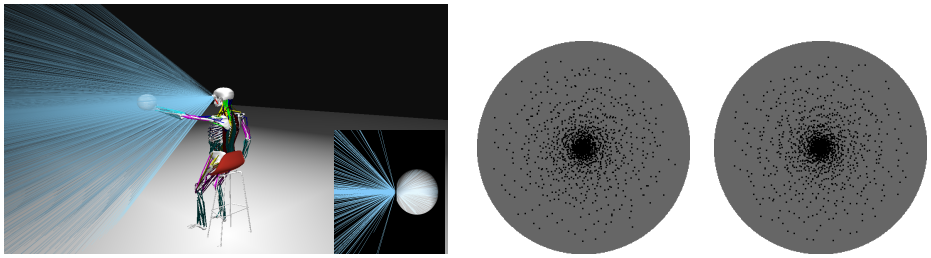
Each arm musculoskeletal complex is rooted at the clavicle and scapula, and includes the humerus, ulnar, and radius bones. A total of 29 Hill-type uniaxial muscles actuate the shoulder, elbow, and wrist joints in each arm. The hands, which currently are actuated kinematically, serve as end effectors.

Each leg musculoskeletal complex, which includes the femur, patella, tibia, and fibula bones, is rooted by the pelvis. A total of 39 Hill-type uniaxial muscles actuate the hip, knee, and ankle joints in each leg. The feet, which currently are actuated kinematically, serve as end effectors.

The motor subsystem (lower box of Fig. 1) includes 10 motor DNNs (numbered 11–20 in the figure) comprising five recurrent neuromuscular motor controllers. Two DNNs control the 216 neck muscles that balance the head atop the cervical column against the downward pull of gravity and actuate the neck-head biomechanical complex, thereby producing controlled head movements. Two DNNs control each limb; in particular, the 29 muscles in each of the two arms and the 39 muscles in each of the two legs. Additional details about our biomechanical human musculoskeletal model and the five recurrent neuromuscular controllers comprising its motor subsystem, are presented elsewhere [5]. The next two sections describe the sensory subsystem, which is illustrated in the upper box of Fig. 1.

3 Ocular and Retinal Models

We modeled the eyes by taking into consideration the physiological data from an average human. As shown in Fig. 1e, we model the virtual eye as a sphere of 12mm radius, that can be rotated with respect to its center around its vertical



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

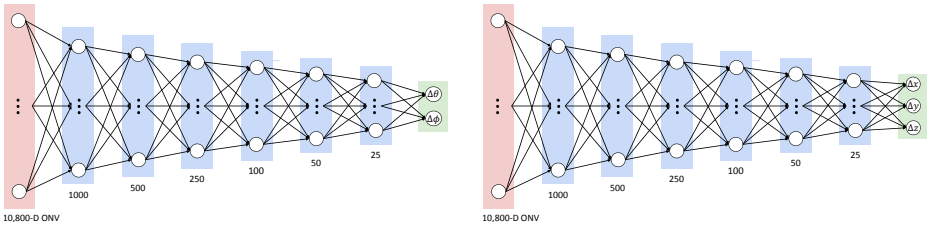
Fig. 2: Via raytracing (blue rays), the biomechanical virtual human's eyes (b) visually sample the environment (a) to compute the irradiance at RGB photoreceptors (black dots) on their foveated retinas (c) (d), which are nonuniformly positioned according to a noisy log-polar distribution.

y axis by a horizontal angle of θ and around its horizontal x axis by a vertical angle of ϕ. The eyes are in their neutral positions looking straight ahead when $\theta = \phi = 0°$. For now, we have modeled the eye as an idealized pinhole camera with aperture at the center of the pupil and with horizontal and vertical fields of view of 167.5°.

We can compute the irradiance at any point on the hemispherical retinal surface at the back of the eye using the conventional raytracing technique of computer graphics rendering [8]. Sample rays from the positions of photoreceptors on the hemispherical retinal surface are cast through the pinhole and out into the 3D virtual world where they recursively intersect with the visible surfaces of virtual objects and query the virtual light sources according to the Phong local illumination model. The irradiance values returned by these rays determine the light impinging upon the retina at the photoreceptor positions. Fig. 2a,b illustrates this retinal imaging process.

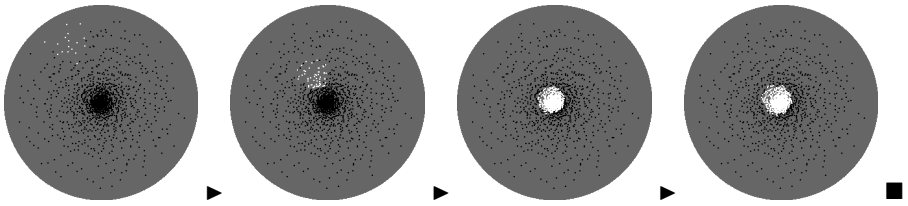
To simulate foveated perception, we use a noisy log-polar distribution to determine the nonuniform 2D positions of the photoreceptors on the retina. Fig. 2c,d illustrates the arrangement of the photoreceptors. By drawing different random numbers, the 3,600 photoreceptors are placed in slightly different positions on each of the two hemispherical retinas. Of course, other placement patterns are possible, including more elaborate biomimetic procedural models or photoreceptor distributions empirically measured from biological eyes, all of which deviate dramatically from the uniformly-sampled Cartesian pixel images commonly used in vision and graphics.

The foveated retinal RGB "image" captured by each eye is output for further processing down the visual pathway, not as a conventional 2D array of pixels, but as a 1D vector of length $3{,}600 \times 3 = 10{,}800$, which we call the Optic Nerve Vector (ONV). The raw sensory information encoded in this vector feeds the vision DNNs that directly control eye movements and whose outputs also feed the motor networks that control head movements and the reaching actions of the limbs.



(a) Foveation DNN (b) Limb vision DNN

Fig. 3: The vision DNN architecture.



(a) t_0 (b) t_1 (c) t_2 (d) t_3

Fig. 4: Time sequence (a)–(d) of photoreceptor responses in the left retina during a saccadic eye movement that foveates and tracks a moving white ball. At time t_0 the ball becomes visible in the periphery, at t_1 the eye movement is bringing the ball towards the fovea, and the moving ball is being fixated in the fovea at times t_2 and t_3.

4 Sensory Subsystem

We next present the 10 vision DNNs (numbered 1–10 in Fig. 1) that implement the sensory subsystem. The sensory subsystem includes two types of fully-connected feedforward DNNs (Fig. 3), which will be described in the next two sections. These DNNs input the sensory information provided by the 10,800-dimensional ONV. The first type (Fig. 3a) controls the eye movements, as well as the head movements via the neck motor DNN. The second type (Fig. 3b) produces arm-to-target 3D error vectors $[\Delta x, \Delta y, \Delta z]^T$ that drive the limbs via the limb motor DNNs.

4.1 Foveation DNNs (1,6)

The first role of the left and right foveation DNNs is to induce saccadic eye movements to foveate visible objects of interest by rotating the eyes to look directly at those visual targets, thereby observing them with maximum visual acuity. Both eyes verge naturally to focus together on a foveated target. This is illustrated in Fig. 4 for a white ball in motion that enters the eye's field of view from the lower right, stimulating several peripheral photoreceptors at the upper left of the retina. The maximum speed of saccadic eye movements is

900 degrees/sec, and the eye almost instantly foveates the visual target. Fine adjustments comparable to microsaccades can be observed during fixation.

The eye movements are tightly coupled with head movements that facilitate foveation, fixation, and visual tracking. By quickly saccading to and then pursuing the visual target, the eyes look directly at it, whereas the head follows, albeit much more sluggishly due to its substantial mass. Hence, the second role of these two DNNs is to control head movement, which is accomplished by driving the neck neuromuscular motor controller (11,12) (Fig. 1f) with the average of their outputs.

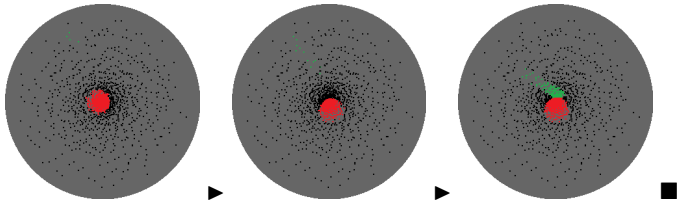
As shown in Fig. 3a, the input layer to the foveation DNN comprises 10,800 units, due to the dimensionality of the ONV, the output layer has 2 units, $\Delta\theta$ and $\Delta\phi$, and there are 6 hidden layers. We conducted a systematic set of experiments with various DNN architectures, activation functions, and other parameters to determine that this architecture was suitable for our purposes [4]. The DNN applies the rectified linear unit (ReLU) activation function, and its 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. We chose Adaptive Moment Estimation (Adam) [1] as the stochastic optimization method, using the following parameters: $lr = 1.0 \times 10^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1.0 \times 10^{-8}$, $\alpha = 0.001$, where β_1 and β_2 represent the exponential decay rate for momentum estimates taking an average and an average of squared gradients, respectively, ϵ prevents a divide-by-zero error, lr is the learning rate, and α is the step size. An early stopping condition is set to avoid overfitting and mean squared error serves as the loss function.

We use our virtual human model to train the network, as follows: We presented a white sphere within the visual field. By raytracing the 3D scene, the photoreceptors in the retinas of each eye are stimulated, and the visual stimuli are presented as the RGB components of the respective ONV. Given this ONV as input, the desired output of the network is the angular differences $\Delta\theta$ and $\Delta\phi$ between the actual gaze directions of the eyes and the known gaze directions that would foveate the sphere. Repeatedly positioning the sphere at random locations in the visual field, we generated a large training dataset of 1M input-output pairs. The backpropagation DNN training process converged to a small error after 80 epochs, which triggered the early stopping condition (no improvement for 10 successive epochs) to avoid overfitting.

4.2 Limb Vision DNNs (2,3,4,5 & 7,8,9,10)

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 to touch the target. This is illustrated in Fig. 5 for a fixated red ball and a (green) arm that enters the eye's field of view from the lower right, stimulating several peripheral photoreceptors at the upper left of the retina.

The architecture of the limb vision DNN (Fig. 3b) is identical to the foveation DNN except for the size of the output layer, which has 3 units, Δx, Δy, and



(a) t_0 (b) t_1 (c) t_2

Fig. 5: Retinal images during an arm reaching motion that deflects a moving red ball. The photoreceptors are simultaneously stimulated by the fixated ball and by the (green) arm entering the eye's field of view from the lower right (upper left on the retina).

Δz, to encode the estimated discrepancy between the 3D positions of the end effector and the visual target.

Again, we use our virtual human model to train the four limb networks, as follows: We present a red ball in the visual field and allow the trained foveation DNNs to foveate the ball. Then, we extend a limb (arm or leg) towards the ball. Again, by continually raytracing the 3D scene, the photoreceptors in the retinas of each eye are stimulated and the visual stimuli are presented as the RGB components of the respective ONV. With the ONV as input, the desired output of the network is the 3D discrepancy, Δx, Δy, and Δz, between the known 3D positions of the end effector and the visual target. Repeatedly placing the sphere at random positions in the visual field and randomly articulating the limb to reach for it in space, we again generated a large training dataset of 1M input-output pairs. The backpropagation DNN training process converged to a small error after 388 epochs, which triggered the early stopping condition to avoid overfitting. As expected, due to the greater complexity of this task, the training speed is significantly slower than that of the foveation DNN.

5 Experimental Results

Fig. 6 shows a sequence of frames from a simulation demonstrating the active sensorimotor system. A cannon shoots a ball towards the virtual human, which actively perceives the ball on its foveated retinas. The ONVs from the eyes are processed by the vision DNNs to enable foveation and visual tracking of the incoming ball. Simultaneously fed by the vision DNNs, the motor DNNs control the extension of the arms and legs to intercept the incoming ball and deflect it out of the way. Thus, given just the high level objective of deflecting the incoming ball, the virtual human successfully controls itself to carry out this nontrivial dynamic sensorimotor control task. A number of additional demonstrations are presented in [5].

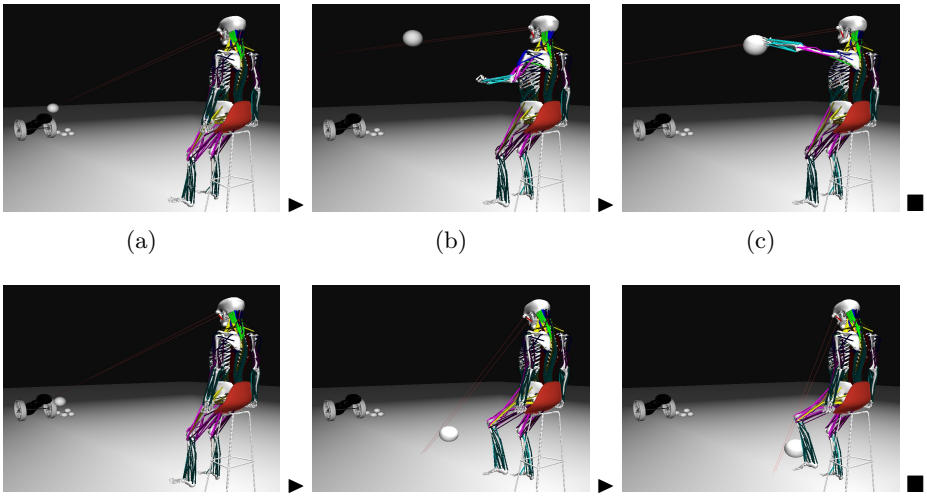


Fig. 6: Frames from a simulation of the biomechanical virtual human sitting on a stool, demonstrating active visual perception and simultaneous motor response; in particular, a left-arm reaching action (a)–(c) and a left-leg kicking action (d)–(f) to intercept balls shot by a cannon. Each incoming ball is perceived by the eyes, processed by the perception DNNs, foveated and tracked through eye movements in conjunction with muscle-actuated head movements controlled by the cervicocephalic neuromuscular motor controller, while visually guided, muscle-actuated limb movements are controlled by the left arm and left leg neuromuscular motor controllers.

6 Conclusion

We have introduced a simulation framework for biomimetic human perception and sensorimotor control. It is unique in that it features an anatomically accurate, biomechanical human musculoskeletal model that is actuated by numerous contractile skeletal muscles. The novel contributions of our framework include the following primary ones:

- A biomimetic, foveated retinal model, which is deployed in a pair of human-like foveated eyes capable of realistic eye movements, that employs raytracing to compute the irradiance captured by a nonuniform distribution of photoreceptors.
- A fully functional prototype sensorimotor system, which in addition to deep-learning-based neuromuscular control of the neck-head, arms, and legs (under the influence of gravity), includes a sensory subsystem that incorporates a set of 10 automatically-trained deep neural networks driven by the optic nerve outputs from the eyes.
- Demonstration of the performance of our innovative sensorimotor system in nontrivial tasks. These simultaneously involve eye movement control for

saccadic foveation and pursuit of a visual target in conjunction with appropriate dynamic head motion control, plus visually-guided dynamic limb control to generate natural limb reaching actions in order to deflect moving visual targets.

Our approach has been to train the deep neural networks with very large quantities of training data that are synthesized by the biomechanical human musculoskeletal model itself. Our work to date confirms that our innovative deep learning approach to strongly biomimetic sensorimotor control works remarkably well. Nevertheless, our prototype visuomotor system inevitably has some limitations that we plan to address in future work. These include appropriately detailed biomechanical modeling of the eye and development of neuromuscular controllers for the 6 extraocular muscles, the incorporation of visual attention and stereoscopic vision mechanisms, which will require a substantial increase in the number of retinal photoreceptors in conjunction with sparsely-connected vision DNN architectures, and ultimately fully unconstraining our virtual human and successfully training thoracic/lumbar neuromuscular controllers, which would enable us to address a variety of more complex sensorimotor tasks.

References

1. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
2. Lee, S.H., Sifakis, E., Terzopoulos, D.: Comprehensive biomechanical modeling and simulation of the upper body. ACM Trans. Graphics **28**(4), 99:1–17 (Aug 2009)
3. Lee, S.H., Terzopoulos, D.: Heads up! Biomechanical modeling and neuromuscular control of the neck. ACM Transactions on Graphics **23**(212), 1188–1198 (2006)
4. Nakada, M., Chen, H., Terzopoulos, D.: Deep learning of biomimetic visual perception for virtual humans. In: ACM Symposium on Applied Perception (SAP 18). pp. 20:1–8. Vancouver, BC (August 2018)
5. Nakada, M., Zhou, T., Chen, H., Weiss, T., Terzopoulos, D.: Deep learning of biomimetic sensorimotor control for biomechanical human animation. ACM Trans. Graphics **37**(4), 56:1–15 (2018), (in Proc. ACM SIGGRAPH 2018)
6. Rabie, T.F., Terzopoulos, D.: Active perception in virtual humans. In: Proc. Vision Interface 2000. pp. 16–22. Montreal, Canada (2000)
7. Schwartz, E.L.: Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. Biological Cybernetics **25**(4), 181–194 (1977)
8. Shirley, P., Morley, R.K.: Realistic Ray Tracing. A. K. Peters, Ltd., Natick, MA, USA, 2 edn. (2003)
9. Terzopoulos, D., Rabie, T.F.: Animat vision: Active vision with artificial animals. In: Proc. Int. Conf. Computer Vision (ICCV). pp. 840–845. Cambridge, MA (1995)
10. Yeo, S.H., Lesmana, M., Neog, D.R., Pai, D.K.: Eyecatch: Simulating visuomotor coordination for object interception. ACM Trans. Graphics (TOG) **31**(4), 42 (2012)