Learning Biomimetic Perception for Human Sensorimotor Control

Masaki Nakada, Honglin Chen, Demetri Terzopoulos
Computer Science Department, University of California, Los Angeles

Abstract

We introduce a biomimetic simulation framework for human perception and sensorimotor control. Our framework features a biomechanically simulated musculoskeletal human model actuated by numerous skeletal muscles, with two human-like eyes whose retinas contain spatially nonuniform distributions of photoreceptors. Its prototype sensorimotor system comprises a set of 20 automatically-trained deep neural networks (DNNs), half of which comprise the neuromuscular motor control subsystem, whereas the other half are devoted to the visual perception subsystem. Directly from the photoreceptor responses, 2 perception DNNs control eye and head movements, while 8 DNNs extract the perceptual information needed to control the arms and legs. Thus, driven exclusively by its egocentric, active visual perception, our virtual human is capable of learning efficient, online visuomotor control of its eyes, head, and four limbs to perform a nontrivial task involving the foveation and visual persuit of a moving target object coupled with visually-guided reaching actions to intercept the incoming target.

1. Introduction

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

Sensorimotor functionality in biological organisms refers to the acquisition and processing of sensory input and the production of appropriate motor output responses to perform desired tasks. In this paper, we introduce a biomimetic simulation framework for investigating human perception and sensorimotor control. Our framework is unique in that it features a biomechanically simulated human musculoskeletal model that currently includes 823 skeletal muscle actuators. Our virtual human actively perceives its environment with two eyes, whose retinas contain photoreceptors arranged in spatially nonuniform distributions.

The prototype visuomotor control system of our human model consists of a set of 20 automatically-trained, fully-connected DNNs that operate continuously and synergistically, half of which comprise the neuromuscular motor control subsystem, while the other half are devoted to the visual perception subsystem, as shown in Fig. 1. Directly from the photoreceptor responses, 2 perception DNNs in the perception subsystem (top half of Fig. 1) control eye and head movements, while 8 DNNs extract the perceptual information needed to control the arms and legs.[1] Thus, driven exclusively by its egocentric, active visual perception, our virtual human is capable of learning efficient, online visuomotor control of its eyes, head, and four limbs to perform nontrivial tasks involving the foveation and visual pursuit of a moving target object coupled with visually-guided reaching actions to intercept the incoming target.

Our prototype 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 through a modular set of DNNs that are automatically trained from data synthesized by the model itself.

2. Related Work

Terzopoulos and Rabie [12] proposed a biomimetic active vision system with foveated perception and visuomotor control for biomechanically-simulated virtual animals. They also applied their "animat vision" visuomotor system to virtual humans, demonstrating vision-guided bipedal locomotion and navigation [9].

In computer graphics, Yeo et al. [14] presented a visuomotor system for an anthropomorphic virtual character capable of visual target estimation tasks and realistic ball catching actions, although their character is purely kinematic rather than biomechanically simulated, and it predicts the trajectories of thrown balls from their known positions and velocities in 3D space without any

[1] In the motor subsystem (bottom half of Fig. 1), two DNNs control the 216 neck muscles that balance the head atop the cervical column against the downward pull of gravity and actuate the cervicocephalic biomechanical complex, thereby producing controlled head movements, and 8 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. See [8] for the details.

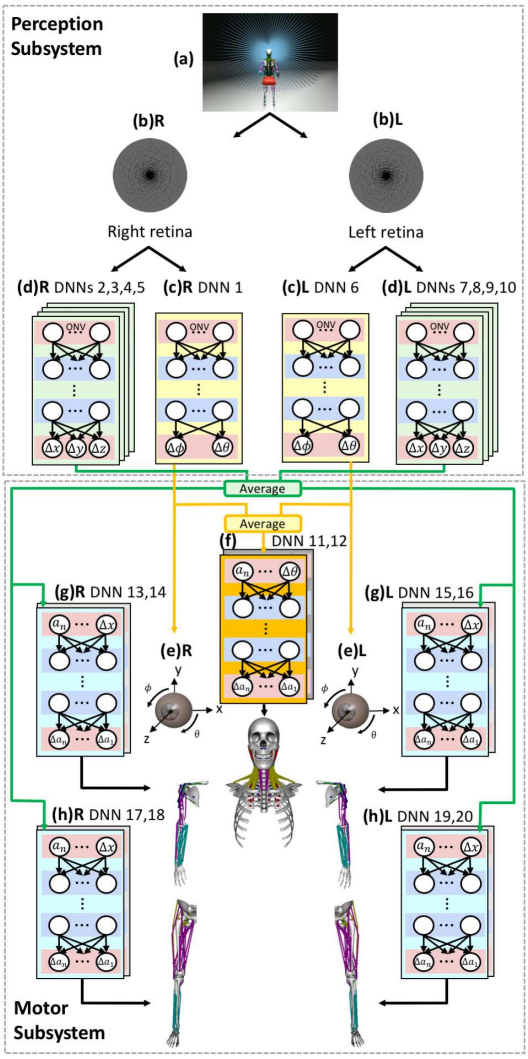


Figure 1: The sensorimotor system architecture, whose controllers include a total of 20 DNNs. In the ***perception subsystem*** (top), each retinal photoreceptor casts a ray into the virtual world (a), which computes and returns the irradiance at that photoreceptor. (b) The arrangement of the photoreceptors (black dots) on the left (b)L and right (b)R foveated retinas. Each eye outputs an Optic Nerve Vector (ONV). (c) The two (yellow) perception DNNs (1,6) input the ONV and produce outputs that drive the movements of the eyes (e) to foveate visual targets. The eight (green) perception DNNs (d)—i.e., (d)L (7,8,9,10) for the left eye and (d)R (2,3,4,5) for the right eye—also input the ONV and output the observed limb-to-target discrepancy estimates. In the ***motor subsystem*** [8] (bottom), the (orange) neck voluntary neuromuscular DNN (f) (11) inputs the average response of the left (c)L and right (c)R foveation DNNs along with the current activations of the 216 neck muscles and produces outputs that actuate the cervicocephalic complex. The four (blue) limb voluntary neuromuscular DNNs (g),(h) (13,15,17,19) input the average response of the left (d)L and right (d)R perception DNNs along with the current activations of the 29 arm or 39 leg muscles and produce outputs that actuate the limbs. The remaining reflex neuromuscular DNNs (f),(g),(h) (12,14,16,18,20) play a stabilizing role [8].

biologically-inspired visual processing. The same is true for the earlier visuomotor system described by Lee and Terzopoulos [7], which was nevertheless incorporated into a biomechanically-simulated model with neuromuscular control not unlike the one described in the present paper.

The virtual animals and humans demonstrated in [12, 9] are equipped with foveated eyes implemented as coaxial virtual cameras capable of eye movements. Using polygon-shaded computer graphics rendering through the GPU pipeline, these virtual eyes capture retinal images as composited multiresolution pyramids supporting foveal and peripheral perception, albeit with a small number of uniformly pixelated pyramidal levels. Our retinal model is significantly more biomimetic. Unlike the uniform, Cartesian grid arrangement of most artificial imaging sensors, visual sampling in the primate retina is known to be strongly space variant [10]. The density of cones decreases radially from the fovea toward the periphery. The log-polar photoreceptor distribution model is commonly used in space-variant image sampling [5, 2, 13]. Given its fundamentally nonuniform distribution of photoreceptors, our virtual retina captures the light intensity in the scene using raytracing [11], which emulates how the human retina samples scene radiance from the incidence of light on its photoreceptors.

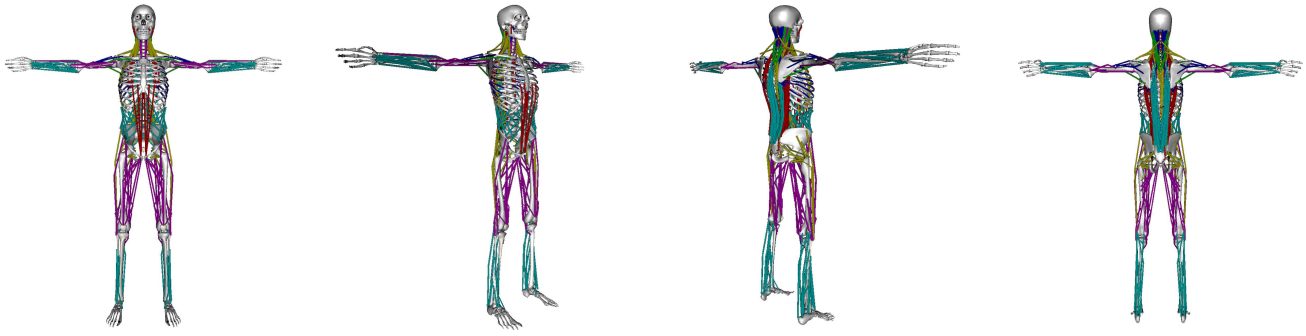
3. Biomechanical Human Model

Fig. 2 shows the musculoskeletal system of our anatomically accurate human model. It includes all of the relevant articular bones and muscles—103 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.[2] 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 (refer to [6] for the details).

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. 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 the cervical column, arms, and legs free to articulate.

Additional details about our biomechanical human musculoskeletal model and the 10 neuromuscular controllers comprising its motor subsystem (see Footnote 1 and the

[2]For the purposes of the research reported in the present paper, the finite element soft-tissue simulation, which produces realistic flesh deformations, is unnecessary and it is excluded to reduce computational cost.



(a) (b) (c) (d)

Figure 2: The biomechanical model, showing the musculoskeletal system with its 103 bones and 823 muscle actuators.

lower half of Fig. 1) are presented elsewhere [8]. The remainder of this short paper develops the perception subsystem, which is illustrated in the top half of Fig. 1.

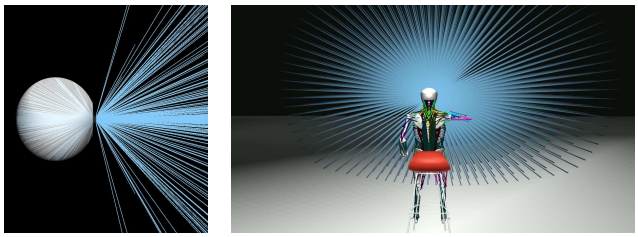
4. Eye and Retina Models

Eye model: We modeled the eyes by taking into consideration the physiological data from an average human.[3] As shown in Fig. 1(e), we model the virtual eye as a sphere of radius of 12mm, that can be rotated with respect to its center around its vertical 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 $\theta = \phi = 0°$. At least for now, we model 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 well-known raytracing technique of computer graphics rendering [11]. Fig. 3 illustrates the retinal "imaging" process. 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 source(s) in accordance with the standard Phong local illumination model. The irradiance values returned by these rays determine the light impinging upon the photoreceptors.

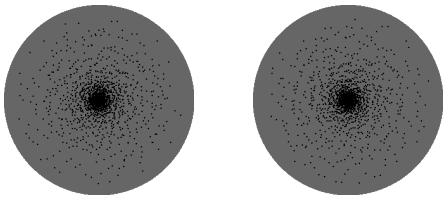
Photoreceptor placement: To simulate biomimetic foveated perception, we position the photoreceptors on the hemispherical retina according to a noisy log-polar distribution. On each retina, we include 3,600 photoreceptors

[3]The transverse size of an average eye is 24.2 mm and its sagittal size is 23.7 mm. The approximate field of view of an individual eye is 30 degrees to superior, 45 degrees to nasal, 70 degrees to inferior, and 100 degrees to temporal. When two eyes are combined, the field of view becomes about 135 degrees vertically and 200 degrees horizontally.



(a) (b)

Figure 3: (a) Rays cast from the positions of photoreceptors on the retina through the pinhole aperture and out into the scene by the raytracing procedure that computes the irradiance responses of the photodectors. (b) All the cast rays as the seated virtual human looks forward with both eyes.



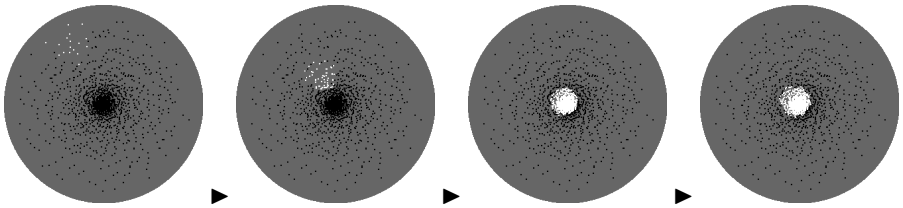
(a) Left retina (b) Right retina

Figure 4: Location of the photoreceptors (black dots) on the left retina (a) and right retina (b) according to the noisy log-polar model.

situated at $\mathbf{d}_k$, for $1 \leq k \leq 3{,}600$, such that

$$\mathbf{d}_k = e^{\rho_j} \begin{bmatrix} \cos\theta_i \\ \sin\theta_i \end{bmatrix} + \begin{bmatrix} \mathcal{N}(\mu, \sigma^2) \\ \mathcal{N}(\mu, \sigma^2) \end{bmatrix}, \qquad (1)$$

where $0 < \rho_j \leq 40$, incremented in steps of 1, and $0 \leq \theta_i < 360°$, incremented in $4°$ steps, and where $\mathcal{N}$ is additive IID Gaussian noise of mean $\mu = 0$ and variance $\sigma^2 = 0.0025$, which places the photoreceptors in slightly different positions on the two retinas. Fig. 4 illustrates the arrangement of the photoreceptors on the left and right reti-



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

Figure 5: 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.

nas. Other placement patterns are readily implementable, including more elaborate procedural models [1] 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.

Optic nerve vectors: The foveated retinal RGB "image" captured by each eye is output for further processing down the visual pathway, not as a 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 perceptions DNNs that directly control eye movements and extract perceptual information that is passed on to the neuromuscular motor control DNNs in the motor subsystem that control head movements and the reaching actions of the limbs.

5. Sensorimotor System

Fig. 1 overviews the sensorimotor system, showing its perception and motor subsystems. The figure caption describes the information flow and the functions of its 20 DNN controllers (labeled 1–20 in the figure). The details of the eyes (Fig. 1(e)) and their retinas (Fig. 1(b)) were presented in the previous section. We will now discuss in greater detail the 10 perception DNNs (labeled 1–10 in Fig. 1).

The perception subsystem includes two types of fully-connected feedforward DNNs that input the sensory information provided by the 10,800-dimensional ONV. The first type controls the eye movements, as well as the head movements via the neck neuromuscular motor controller. The second type produces arm-to-target 3D error vectors $[\Delta x, \Delta y, \Delta z]^T$ that drive the limbs via the limb neuromuscular motor controllers. Both types are described in the next two sections.

5.1. Foveation DNNs (1,6)

The first role of the left and right foveation DNNs is to generate changes in the gaze directions that drive saccadic eye movements to foveate visible objects of interest, thereby observing them with maximum visual acuity, as is illustrated in Fig. 5 for a white ball in motion that enters the field of view from the lower right, stimulating several peripheral photoreceptors in the upper left peripheral region of the retina. The eye almost instantly performs a saccadic rotation to foveate the visual target. Fine adjustments comparable to microsaccades are observed during fixation.

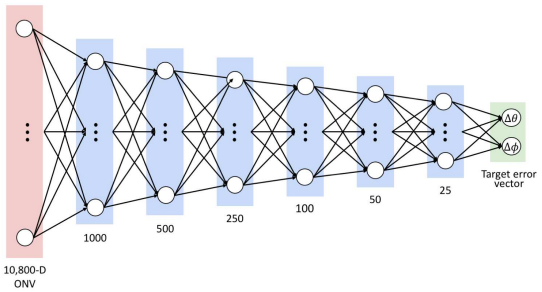
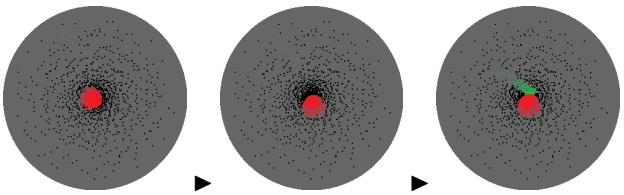


Figure 6: The fully-connected feedforward perception neural network architecture. The network shown is for foveation eye movements.

The second role of these two DNNs is to control head movement, which is accomplished simply by driving, with the average of their outputs, the aforementioned neck neuromuscular motor controller (DNNs 11,12) (Fig. 1(f)). The kinematic eye movements are tightly coupled with the dynamic head movements that facilitate fixation and visual tracking.

Network architecture: As Fig. 6 shows, the input layer to this DNN comprises 10,800 units, to accommodate the dimensionality of the ONV, the output layer has 2 units, $\Delta\theta$ and $\Delta\phi$, and there are 6 hidden layers.[4] The DNN uses 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 [3]. We employ the mean-squared-error loss function and the Adaptive Moment Estimation (Adam) [4] stochastic optimizer with

[4] We conducted experiments with various DNN architectures, activation functions, and other parameters to determine suitable architectures for our purposes.



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

Figure 7: Retinal images during an arm reaching motion that deflects a moving ball. The photoreceptors are simultaneously stimulated by the fixated red ball and by the green arm entering the eye's field of view from the lower right (upper left on the retina).

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, and overfitting is avoided using an early stopping condition.

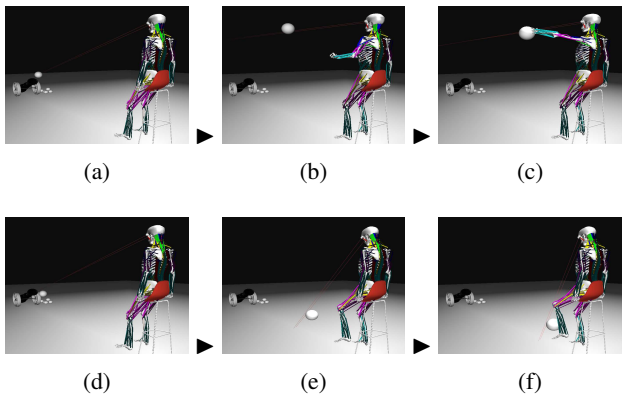
Training data synthesis and network training: We use our virtual human model to train the network, as follows: We presented a white sphere within the visual field. By raytracing the 3D scene, the photoreceptors in the retinas of each eye are stimulated, and the visual stimuli are presented as the RGB 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 an early stopping condition (no improvement for 10 successive epochs) to avoid overfitting.

5.2. Limb Perception DNNs (2,3,4,5 & 7,8,9,10)

The role of the left and right limb (arm and leg) perception 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. 7 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.

Network architecture: The architecture of the limb perception DNN is identical to the foveation DNN in Fig. 6, except for the size of the output layer, which has 3 units, Δx, Δy, and Δz, to encode the estimated discrepancy between the 3D positions of the end effector and the visual target.

Data synthesis and training: Again, we use our virtual human model to train the four limb networks, as follows:



(a) (b) (c)

(d) (e) (f)

Figure 8: 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.

We present a red ball in the visual field and have the trained foveation DNNs foveate the ball. Then, we extend a limb (arm or leg) towards the ball. Again, 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 its 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 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.

6. Experimental Results

Fig. 8 shows a sequence of frames from a simulation demonstrating the active sensorimotor system. A cannon shoots balls towards the virtual human, which it actively perceives with its eyes and reaches out with its arms and legs to intercept. Its 20 DNNs operate continuously and synergistically. The ONVs from the retinas are processed by the pair of foveation DNNs, enabling the foveation and visual tracking of the incoming balls through eye move-

ments coupled with cooperative head movements that follow the gaze direction. The head movements are actuated by the neuromuscular cervicocephalic motor controller, which is fed by the average of the foveation DNN outputs. Naturally, the head movements are much more sluggish than the eye movements due to the considerable mass of the head. Simultaneously, visually-guided by the outputs from the four pairs of limb perception DNNs, the neuromuscular limb motor controllers actuate the arms and legs such that they extend to intercept the incoming balls, deflecting them out of the way. Thus, the biomechanical musculoskeletal human model continuously controls itself to carry out this nontrivial sensorimotor task in an online, (virtual) real-time manner, and no balls shot at it are missed.

7. Conclusion

We have introduced a simulation framework for investigating biomimetic human perception and sensorimotor control. Our framework is unique in that it features an anatomically accurate, biomechanically simulated virtual human model that is actuated by numerous contractile skeletal muscles. Our contributions in this paper include the following primary ones:

1. The development of a biomimetic, foveated retina model, which is deployed in a pair of human-like eyes capable of realistic eye movements, that employs ray-tracing to compute the irradiance captured by a multitude of nonuniformly arranged photoreceptors.

2. Demonstration of the performance of our sensorimotor system in tasks that simultaneously involve eye movement control for saccadic foveation and pursuit of visual targets in conjunction with appropriate dynamic head motion control, plus visually-guided dynamic limb control to produce natural arm and leg extension actions that enable the virtual human to intercept the moving target objects.

7.1. Future Work

Our current eye models are idealized pinhole cameras. We plan to create a more realistic model of the eye that includes a finite-aperture pupil capable of dilation and constriction to control the incoming light, as well as a model of the lens of the eye that would refract the cast rays passing through it and, via active lens deformation, be capable of focusing the image onto the retina, thus synthesizing depth of field phenomena.

Our current eye models are also purely kinematically rotating spheres. We plan to implement a fully dynamic eye model in which the sphere has the typical 7.5 gram mass of the human eyeball and is actuated by the set of 6 extraocular muscles, including the 4 rectus muscles that actuate much of the θ, ϕ movement of our kinematic eyeball, but also the 2 oblique muscles that induce torsion in the gaze direction, around the eye's z axis.

Our vision system generates saccadic eye movements to foveate interesting objects in a variety of different scenarios. Hence, our model can be valuable in human visual attention research, a topic that we wish to explore in future work.

The jobs of the DNNs that must estimate from their ONV inputs the discrepancy between the 3D positions of the end effector and visual target are made difficult by the fact that 3D depth information is lost with projection onto the 2D retina and, in fact, the estimation of depth discrepancy is currently quite poor. This limitation provides an opportunity to explore binocular stereopsis with an enhanced version of our foveated perception model. For this, as well as for other types of subsequent visual processing, we will likely want to increase the number of photoreceptors, experiment with different nonuniform photoreceptor organizations, and automatically construct 2D retinotopic maps from the 1D ONV inputs.

References

[1] M. F. Deering. A photon accurate model of the human eye. *ACM Trans. Graphics*, 24(3):649–658, 2005. 4

[2] L. Grady. *Space-variant computer vision: A graph-theoretic approach*. PhD thesis, Boston University, 2004. 2

[3] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proc. ICCV*, pages 1026–1034, 2015. 4

[4] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[5] J. Koenderink and A. Van Doorn. Visual detection of spatial contrast. *Biological Cybernetics*, 30(3):157–167, 1978. 2

[6] S.-H. Lee, E. Sifakis, and D. Terzopoulos. Comprehensive biomechanical modeling and simulation of the upper body. *ACM Trans. Graphics*, 28(4):99:1–17, Aug. 2009. 2

[7] S.-H. Lee and D. Terzopoulos. Heads up! Biomechanical modeling and neuromuscular control of the neck. *ACM Trans. Graphics*, 23(212):1188–1198, 2006. 2

[8] M. Nakada, T. Zhou, H. Chen, T. Weiss, and D. Terzopoulos. Deep learning of biomimetic sensorimotor control for biomechanical human animation. *ACM Trans. Graphics*, 37(4):1–14, 2018. Proc. ACM SIGGRAPH 2018. 1, 2, 3

[9] T. F. Rabie and D. Terzopoulos. Active perception in virtual humans. In *Proc. Vision Interface*, pages 16–22, 2000. 1, 2

[10] E. L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194, 1977. 2

[11] P. Shirley and R. K. Morley. *Realistic Ray Tracing*. A. K. Peters, Ltd., Natick, MA, USA, 2 edition, 2003. 2, 3

[12] D. Terzopoulos and T. F. Rabie. Animat vision: Active vision with artificial animals. In *Proc. ICCV*, pages 840–845, 1995. 1, 2

[13] S. W. Wilson. On the retino-cortical mapping. *International Journal of Man-Machine Studies*, 18(4):361–389, 1983. 2

[14] S. H. Yeo, M. Lesmana, D. R. Neog, and D. K. Pai. Eyecatch: Simulating visuomotor coordination for object interception. *ACM Trans. Graphics*, 31(4):42, 2012. 1