Humanoid robots are retrieving emotion from motion analysis

Humanoid Robots are Retrieving Emotion from Motion Analysis

Tino Lourens Emilia Barakova

Eindhoven University of Technology, P.O. Box 513, Eindhoven, The Netherlands

Abstract

This paper presents an application for hand waving in real time using a parallel framework. Analysis of 15 different video fragments demonstrates that acceleration and frequency are relevant parameters for emotion classification of hand waving. Its solution will be used for human-robot interaction with the aim of training autistic children social behavioral skills in a natural environment.

1 Introduction

A little more than a decade ago, Honda Corporation demonstrated an Advanced Step In MObility (ASIMO) [9, 10]. The introduction of ASIMO has led to a boost in humanoid robotic research and development. One of the trends in humanoid robot development is the androids/actroids like Repliee [20] and DER, while another trend are relatively inexpensive open humanoid platforms like Speecys SPC-101C and the NAO humanoid from Aldebarn. Latter has become a standard platform in the robocup soccer competition. Due to their standardized hardware and software, which will make the experiments reproducible, these robots have become an attractive platform for conduction of behavioral studies.

Sociable humanoid robots pose a dramatic and intriguing shift in the way one thinks about control of autonomous robots, and are the first generation of robots where a substantial human-robot interaction is expected [5]. The introduction of mobile robots that have to demonstrate a certain degree of autonomy yield different requirements than an industrial robot or the way ASIMO has been pre-programmed in a conditioned environment. These robots need to have a higher level perceptual, behavior, emotion system, see [5, 2], but also a mechanism to cooperate with uncertainty and one for survival to guarantee a degree of autonomy. In many ways such a system resembles aspects of brain like functional behavior, its evident that such a robot should be able to process information in real time in a highly parallel way.

Recent developments by the personal computers provide a substantial computing power of graphical processing units (GPUs).[1] Moreover the GPUs reach this performance due to massive parallelism making them as powerful as a fast supercomputer of just three years ago. Parallelism is an important aspect for both functional brain modeling [18] and graphical software development [14] and it implies that real world applications that are bound by real time constraints have become feasible.

Our goal is to construct a computational framework for social interaction where both technologies for parallelism are used. More specifically we are interested in scenarios involving multiple simultaneous actions performed by different body parts of a human or a robot. We assume realistic imitation scenarios, i.e., scenarios where a human freely behaves and a robot tracks its actions with the intend to act upon the meaningful ones, for instance by imitation [3] or by encoding episodes [1].

The paper is organized as follows: In Section 2 we focus on the physical characteristics of the used platform and elaborate on the parallel architecture. Section 3 describes the experimental setup and gives the implementation of marking a moving hand in an image using skin color and motion characteristics. For the sequence of images such region is marked to construct a stream that is used to analyze hand waving behavior. Section 4 provides insight how these data streams are used to extract social behavior for interaction with a robot. The paper finishes with a discussion and future research.

[1] A single GPU card is able to process more than one tera (10^{12}) floating point operations per second (TFLOPS).

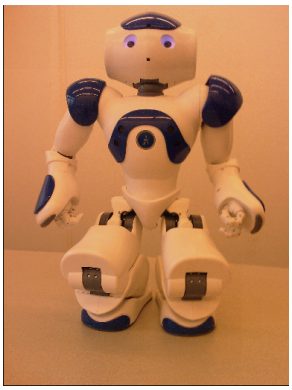

Figure 1: Application platform humanoid robot NAO.

2 Infrastructure

2.1 Humanoid robot

Commercially available humanoid robot NAO, illustrated in Figure 1, is used for the experiment. The robot has 25 degrees of freedom, 5 in each leg and arm, and 1 in each hand. Further it has 2 degrees of freedom in its head and one in the pelvis. The platform contains 2 color cameras with a maximum resolution of 640x480 pixels at a speed of 30 frames per second. The platform contains an embedded AMD Geode 500MHz processor and is shipped with an embedded Linux distribution. A software library called NaoQi is used to control the robot. This API provides an easy to use C++ interface to the robot's sensors and actuators. Due to this library its is relatively easy to control the robots actuators and make use of advanced routines that let the robot move and talk using text to speech.

2.2 Parallel processing units

Graphics processing units (GPUs) were originally designed to perform highly parallel computations required for graphics rendering. But over the last couple of years, they have proven to be powerful computing workhorses across more than just graphics applications. Designed with more resources devoted to data processing rather than flow control and data caching, GPUs can be leveraged to significantly accelerate portions of codes traditionally run on CPUs.

Compute unified device architecture (CUDA) is Nvidia's parallel computing architecture, which manages computation on the GPU in a way that provides a simple interface for the programmer. The CUDA architecture can be programmed in C with a few extensions [8].

2.3 Software

The proposed parallel processing framework assumes a set of useful functional units. These units are connected with each other to exchange information. Figure 2a illustrates such a parallel framework, where processing units and connections are represented by squares and directed arrows, respectively. Using a graphical setup gives fast insight in the available parallel processes and their respective connections, and can be implemented in graphical programming environment TiViPE (www.tivipe.com) preserving its graphical representation [14]. Such network can be transferred into a formal language and described in a BackusNaur Form (BNF), as illustrated Figure 2b. This makes such a framework attractive for TiViPE to generate and compile code fully automatically from a constructed graphical network.

The setup of such a framework is simple and elegant, but powerful enough to describe any parallel algorithm. In this respect it can be used as an artificial neural network, where the squares and connections denote the neurons, and synapses, respectively. The framework could be used as probabilistic network where connections denote the chance to go from one state to another one.

Figure 2c shows a conceptual network of brain areas involved in hand object interaction. In this example the visual input has been described as a single block, but it contains many parallel processing units, as provided in earlier work [22, 17, 15, 16]. Our goal is to make a brain model from a functional perspective using this parallel framework [18].

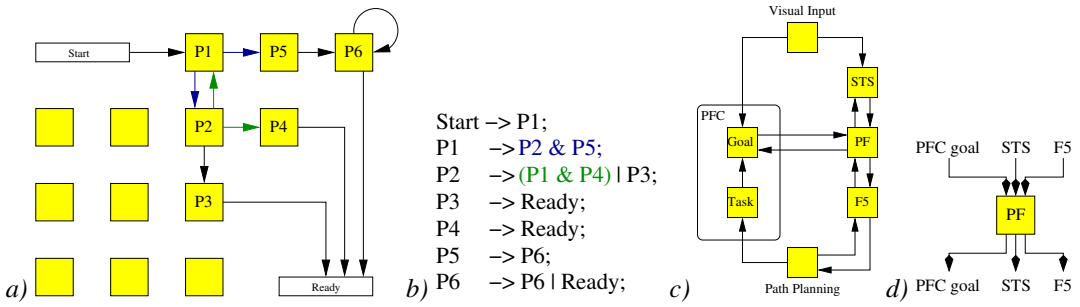

Figure 2: General parallel framework where yellow squares denote processing units. Arrowed connection denote information exchange in direction of the arrow between processing units. *a*) An example of such a network, and *b*) Its textual description. *c*) Brain areas involved in hand object interaction. *d*) Isolated processing unit with input and output connections.

3 Experimental setup

The experiment we have been conducting is hand waving. Figure 5 depicts four (happy, angry, sad, and polite) different emotional waving patterns. A camera records images that are processed using a combination of skin color and motion detection, with the aim of tracking a single area. This area is associated with the waving movement in the most natural way. The robot should be able to extract a simple motion pattern and derive its behavioral aspects, and should be able to imitate this pattern and eventually adjust its behavior, both with the aim of either to teach or to influence behavior of the human to improve his or her social skills.

The implementation of detection and tracking a waving hand is given in Figure 3, and has been decomposed into several building blocks:

1. acquiring data from a camera or reading a stored image sequence
2. binarizing an image by marking a pixel either as skin color or other color and in parallel binarizing an image by marking pixels either as observed motion or as static element
3. marking skin and motion regions by a pyramid decomposition
4. selection of these regions that are both skin and motion region
5. averaging skin-in-motion regions to a single region
6. tracking an averaged skin-in-motion region
7. visualization of regions in an image
8. visualization of a waving hand
9. classification of waving profiles

3.1 Skin color detection

An image is defined as a two-dimensional matrix where a pixel at position (x, y) has an intensity $I_c(x, y) = (r, g, b)$, where r, g, and $b \in [0, \ldots, 255]$ are the red, green, and blue component. Segmentation using skin color can be made independent of differences in race when processing image pixels in Y-Cr-Cb color space [6]. The following (r, g, b) to (Y, Cr, Cb) conversion is used

$$Y = 0.2989r + 0.5866g + 0.1145b; \qquad Cr = 0.7132(r - Y); \qquad Cb = 0.5647(b - Y).$$

Threshold values as used by Chai and Ngan[6]

$$77 < Cb < 127; \qquad 133 < Cr < 173$$

yield good results for classifying pixels belonging to the class of skin tones.

In our experiments we also excluded the "white area". Formally an element belongs to the "white area" if it satisfies the following:

$$\frac{|r - g|}{m} < 0.1 \wedge \frac{|r - b|}{m} < 0.1 \wedge \frac{|g - b|}{m} < 0.1, \tag{1}$$

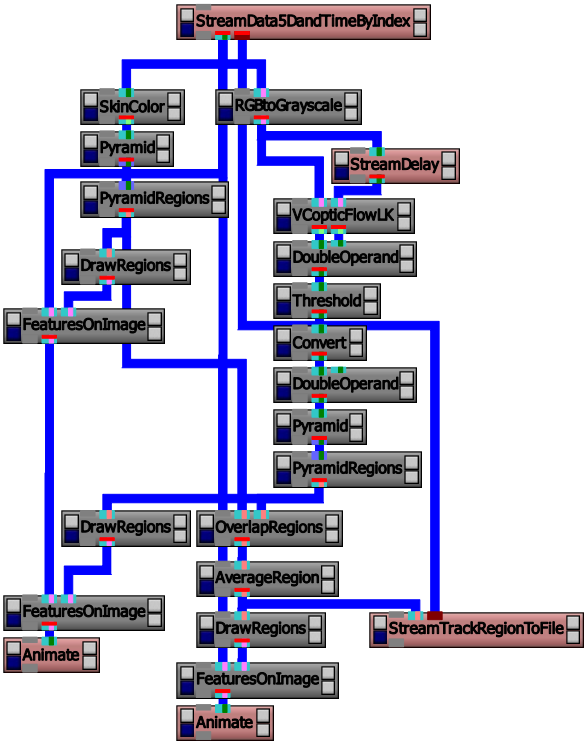

Figure 3: TiViPE [14] implementation of handwaving. The icons from top to bottom at the left-side process skin areas, while motion sensitivity is processed at the right-side.

where $m = \min(r, g, b)$, $r > 0.3$, $g > 0.3$, and $b > 0.3$. Its implementation given by 'the "SkinColor" icon in Figure 3 yields a binary image.

It is plausible that the functional concept as described above contains similarities with how the brain processes visual data, since primary visual cortex area V4 provides a substantial role in processing color [23].

3.2 Motion detection

A pixel at location (x, y) with intensity $I(x, y) = (r + g + b)/3$ will have moved by $(\delta x, \delta y)$ over a time span δt, hence the following image constraint equation can be given:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t). \tag{2}$$

In this method the following needs to be solved:

$$I_x V_x + I_y V_Y = -I_t, \tag{3}$$

where (V_X, V_y) denotes the flow, I_x and I_y the derivatives in horizontal and vertical direction, respectively. The Lucas-Kanade operator [19] is used, it is a two-frame differential method for optical flow estimation where the derivatives are obtained by using a Sobel filter kernel [21]. Instead of using Sobel kernels the more biologically plausible Gabor kernels can be used as well [7, 13]. Receptive fields of a cat's primary visual cortex area V1 and V2, that show striking similarities with these Gabor kernels, have been found in the 1950's by neuroscientists and Nobel laureates Hubel and Wiesel [11, 12]. It is plausible that a similar activity flow map might be expected in the middle temporal are MT, also known as primary visual cortex area V5 [23].

From (V_x, V_y) the L-2 norm (top right "DoubleOperand" icon of Figure 3) is taken and thresholded at 5 to obtain a binary "motion classified" image.

3.3 Rectangular region marking

The next stage is marking a cluster of "skin tone classified" or "motion classified" pixels by a rectangular window. This is performed by decomposing the image into a pyramid, where every pixel in the next level of

Figure 4: Marked regions of interest. Red and green areas denote skin and motion area, respectively. A blue area is the combined moving skin area that has been averaged over all skin areas that are moving.

the pyramid is computed as follows:

$$I_{i+1}(x,y) = \left(I_i(2x,2y) + I_i(2x+1,2y) + I_i(2x,2y+1) + I_i(2x+1,2y+1)\right)/4, \tag{4}$$

where (x,y) is the position in image I_i, i denotes the level in the pyramid, and base level 0 contains the original image I_0. The construction of a pyramid using (4) provides a strongly reduced search space, since if in level $i+1$ a pixel $I_{i+1}(x,y)$ is found to belong to the desired region then in level i of the pyramid a cluster of 2x2 pixels ($I_i(2x,2y)$, $I_i(2x+1,2y)$, $I_i(2x,2y+1)$, and $I_i(2x+1,2y+1)$) belong to the same region.

The search for regions of interest starts at the highest level, and decreases until an a-priori known minimum level has been reached. It is therefore possible that no regions of interest are found. Taking into consideration that if a pixel is marked as "skin tone" or motion it has value 1, and 0 otherwise. We define a pixel to belong to a unique region j if it satisfies the following:

$$R_i^j(x,x+1,y,y+1) = I_i(x,y) \equiv 1. \tag{5}$$

Regions R_i^j in their initial setting are bound by a single pixel $I_i(x,y)$, and a region growing algorithm is applied to determine the proper size of the rectangular region. Lets assume that the initial size of the rectangle is $R_i^j(x_l,x_r,y_u,y_d)$ and that the possible growing areas are left ($R_i^{j^l} = R_i^j(x_l-1,x_l,y_u,y_d)$), right ($R_i^{j^r} = R_i^j(x_r,x_r+1,y_u,y_d)$), above ($R_i^{j^u} = R_i^j(x_l,x_r,y_u-1,y_u)$, and below ($R_i^{j^d} = R_i^j(x_l,x_r,y_d,y_d+1)$ this region. The average value of all four growing areas is taken, where the maximum value determines the direction of growing. The following procedure

$$A_i^{j^x} = \text{avg}\left(R_i^{j^x}\right) \quad x \in \{l,r,u,d\}; \qquad M_i^{j^x} = \max_x\left(A_i^{j^x}\right); \qquad R_i^j = R_i^j \cup R_i^{j^x} \quad \text{if } M_i^{j^x} \geq T_{rg}$$

is repeated until $M_i^{j^x} < T_{rg}$. From experiments T_{rg} = 0.67 provides a rectangle that corresponds roughly to a skin area in the original image and 0.5 gives a sufficiently large motion area, see also Figure 4.

The method described above is able to find all uniform skin color and motion regions in an image in real time.

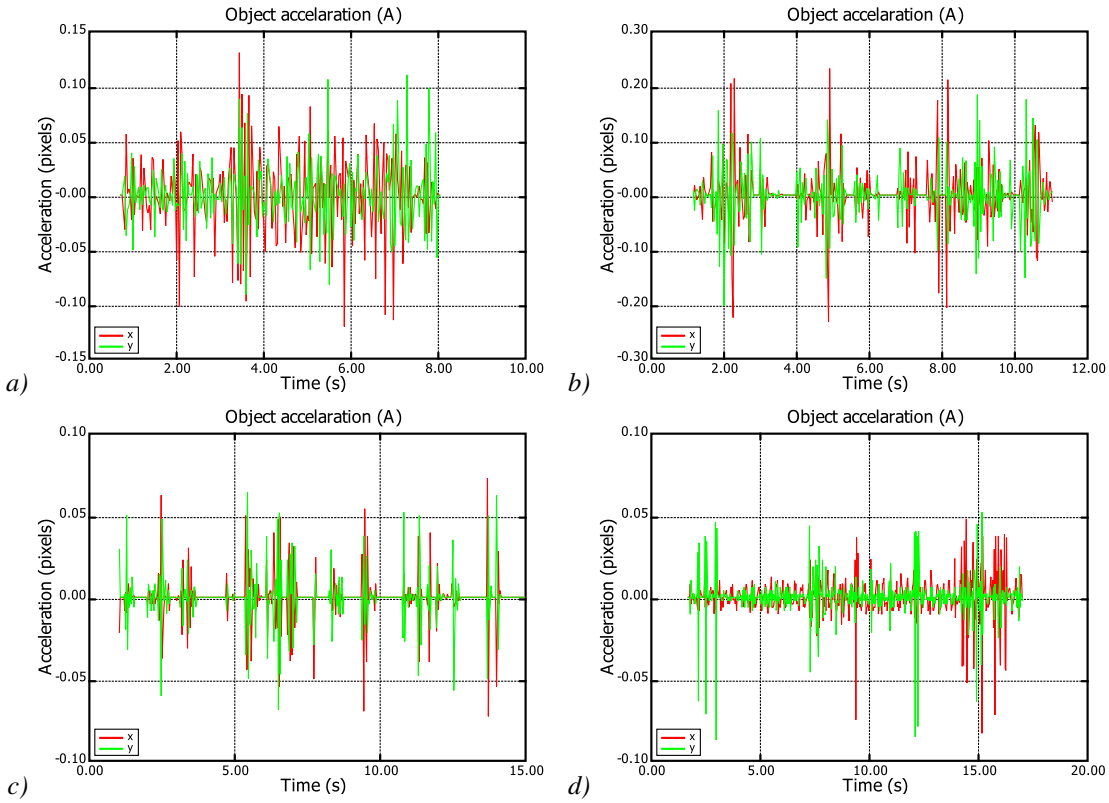

Figure 5: Waving patterns. a-d) show acceleration profiles for happiness, anger, sadness, and politeness, respectively.

Formally such a feature f can be described by its region, type, and time: $f(xl, xr, yu, yd, regiontype, t)$. This f in turn could be further processed by other visual areas or passed on to both STS and PFC, as illustrated in Figure 2c.

3.4 Tracking

Two examples of the waving experiment using color images of 640x480 pixels at a speed of 29 frames per second are provided in Figure 4. The aim of creating a single region in the image rather than multiple regions of interest is being able to unambiguously track an object of interest. These tracked objects are stored as file, see also icon "StreamTrackRegionToFile" of Figure 3, and processed further.

Fifteen recordings of 20 seconds have been made where a Laban expert was asked to demonstrate happiness, anger, sadness, or politeness. In every row the acceleration profile is given it has been obtained by taking the second derivative of the central point of the tracked object. These four different mental states are given in Figure 5, and from this figure the following can be observed:

1. *happy* waving provides a regular waving pattern with a relatively high frequency.
2. *anger* demonstrates bursts with tremendous acceleration
3. *sadness* demonstrates a profile of low acceleration, its frequency is relatively low and appears to have a lower frequency compared to the other three emotions.
4. *politeness* that demonstrates a queen type of waving profile is a regular pattern with a relatively high frequency that is obtained by using minimal energy.

Figure 6 supports these observations. In a an average acceleration-frequency matrix four distinctive clusters are formed. In one of the image sequences the Laban expert was instructed to perform polite waving, but in the sequence she seemed to be happy, indicating that there might be a smooth boundary between these emotional states. The average energy in one of the five bursts of Figure 5c gave an average acceleration score of more than 0.07 and gives an indication of the energetic upper bound.

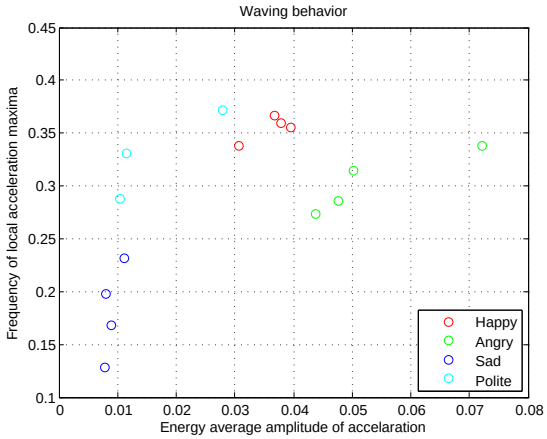

Figure 6: Distinct emotion profiles are revealed by average frequency and acceleration.

4 Behavioral Primitives

Understanding motion from waving patterns requires a mechanism that is able to learn to interpret and classify these sequences, and ideally able to extract the observations provided in Section 3.4. In a complementary study we are attempting to classify motion by so-called Laban primitives. Using these primitives we classify whether the wave pattern demonstrates normal, polite, stressed behavior, or abnormal behavior.

The current method is developed to enable the robot to interact with a human in realistic scenarios. Being able to track regions of interest in parallel, a considerable number of interesting scenarios are possible without have any notion of the meaning of an object. Moreover, using an earlier developed technique [4] the robot recognizes and learns repeating patterns of behavior, which it considers important, and discards occasional movements which most often are not important. For instance, if during waving of the hand a head movement takes place because at this time somebody enters the room, this movement will be ignored. However, if an autistic child performs repeatedly a movement with his/her head while waving, this will be learned and eventually included in the imitation behavior of the robot.

5 Discussion and Future work

We have presented a computational framework that is simple and elegant, but powerful enough to describe any parallel algorithm. This framework is used as basis for our social interaction with the robot, and application of hand waving as well. The goal of this application is to demonstrate some simple social interaction between man and machine.

In this paper we have shown that the robot is able to detect multiple regions in real time and that a stream of such information provides clear insight in the movements of a waving hand. In a complementary study these regions are transfered into behavioral primitives. These primitives are used by the robot to socially interact with a human.

It is obvious that we have barely touched the surface of all the research we would like to conduct. Even a simple experiment like hand waving elicits a number of questions:

- Could any type of waving be predicted?
- How to respond to a waving pattern?
- Does it lead to adaptive or predictive behavior?
- How does the design of simple reactive behavior look like?
- How to design imitation, such that it appears natural and distinctive on a humanoid robot?

Nevertheless, the newly acquired NAO humanoid robots provide an excellent test-bed for machine-interaction, and opens a wide range of possible research areas. An important aspect on these robots will be how closely one can emulate human movement. It implies that understanding the physical limitations of these robots, and getting insight in the parameters settings of the 25 joints of these robots will play an important role for social interaction. Next a basic set of motion primitives needs to be derived that yield the back-bone for social interaction on the basis of body language.

References

[1] E. I. Barakova and T. Lourens. Efficient episode encoding for spatial navigation. *International Journal of Systems Science*, 36(14):877–885, November 2005.

[2] E. I. Barakova and T. Lourens. Analyzing and modeling emotional movements: a framework for interactive games with robots. *Personal and Ubiquitous Computing*, 2009. In press.

[3] E. I. Barakova and T. Lourens. Mirror neuron framework yields representations for robot interaction. *Neurocomputing*, 72(4-6):895–900, 2009.

[4] E.I. Barakova and D. Vanderelst. From spreading of behavior to dyadic interaction -a robot learns what to imitate. *International Journal of Intelligent Systems*, 2009. In press.

[5] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59:119–155, 2003.

[6] D. Chai and K. N. Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564, June 1999.

[7] John G. Daugman. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, July 1988.

[8] T. R. Halfhill. Parallel processing with cuda. *IN-STAT Microprocessor Report*, pages 1–8, January 2008.

[9] K. Hirai. Current and future perspective of Honda humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robotics and Systems*, pages 500–508, 1997.

[10] K. Hirai, M. Hirose, Y. Haikawa, and T. Takenake. The development of Honda humanoid robot. In *IEEE International Conference of Robotics and Automation*, pages 1321–1326, 1998.

[11] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148:574–591, 1959.

[12] D. H. Hubel and T. N. Wiesel. Ferrier Lecture functional architecture of macaque visual cortex. *Proc. R. Soc. Lond. B.*, 198:1–59, July 1977.

[13] T. Lourens. *A Biologically Plausible Model for Corner-based Object Recognition from Color Images*. Shaker Publishing B.V., Maastricht, The Netherlands, March 1998.

[14] T. Lourens. Tivipe –tino's visual programming environment. In *The 28^{th} Annual International Computer Software & Applications Conference, IEEE COMPSAC 2004*, pages 10–15, 2004.

[15] T. Lourens and E. I. Barakova. Tivipe simulation of a cortical crossing cell model. In J. Cabastany, A. Prieto, and D. F. Sandoval, editors, *IWANN 2005*, number 3512 in Lecture Notes in Computer Science, pages 122–129, Barcelona, Spain, June 2005. Springer-verlag.

[16] T. Lourens and E. I. Barakova. Orientation contrast sensitive cells in primate v1 –a computational model. *Natural Computing*, 6(3):241–252, September 2007.

[17] T. Lourens, E. I. Barakova, H. G. Okuno, and H. Tsujino. A computational model of monkey cortical grating cells. *Biological Cybernetics*, 92(1):61–70, January 2005. DOI: 10.1007/s00422-004-0522-2.

[18] T. Lourens, E. I. Barakova, and H. Tsujino. Interacting modalities through functional brain modeling. In J. Mira and J. R. Álvarez, editors, *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks, IWANN 2003*, volume 2686 of *Lecture Notes in Computer Science*, pages 102–109, Menorca, Spain, June 2003. Springer-Verlag.

[19] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging understanding workshop*, pages 121–130, 1981.

[20] K. F. MacDorman and H. Ishiguro. The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies*, 7(3):297–337, 2006.

[21] I. E. Sobel. *Camera Models and Machine Perception*. PhD thesis, Electrical Engineering Department, Stanford University, Stanford, CA, 1970.

[22] R. P. Würtz and T. Lourens. Corner detection in color images through a multiscale combination of end-stopped cortical cells. *Image and Vision Computing*, 18(6-7):531–541, April 2000.

[23] S. Zeki. *A Vision of the Brain*. Blackwell science Ltd., London, 1993.