

Retrieving Emotion from Motion Analysis

in a real time parallel framework for robots

Tino Lourens and Emilia Barakova

Eindhoven University of Technology
P.O. Box 513, Eindhoven, The Netherlands
`t.lourens@tue.nl` and `e.i.barakova@tue.nl`

Abstract This paper presents a parallel real time framework for emotion extraction from video fragments of human movements. Its framework is used for tracking of a waving hand by evaluation of moving skin-colored objects. The tracking analysis demonstrates that acceleration and frequency characteristics of the traced objects are relevant for classification of the emotional expressiveness of human movements. The solution is part of a larger project on interaction between a human and a humanoid robot with the aim of training social behavioral skills to autistic children with robots acting in natural environment.

1 Introduction

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. The introduction of mobile robots that have to demonstrate a certain degree of constitutive autonomy yield different requirements than industrial robots that have been pre-programmed to work in a fully controlled environment. Sociable robots need to have even higher level of autonomy, dealing not only with its perceptual, and behavioral aspects, but also with its interactive aspects, as present for instance in the emotion system [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. We have adopted the approach of functional brain modeling [16] and use graphical software environment TiViPE [12] to realize and integrate these functional models, in a similar way as in earlier work [19,15,13,14]. Real time parallelism on a PC has been strongly facilitated by recent developments of graphical processing units (GPUs), not only have these GPUs become fast¹ but they also can be used as general processing units [8].

We are interested in scenarios involving multiple simultaneous actions performed by different body parts of a human or a robot. We assume realistic

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

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]. In this paper the focus is on hand waving with the aim of detecting different emotional states that can be used either to imitate or to influence the emotional state.

The paper is organized as follows: Section 2 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 3 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.

2 Experimental setup

We have been conducting hand waving experiments within scenarios where different emotions has been enacted. waving. Figure 3 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. A device or robot should be able to extract a simple motion pattern and interpret the intend or the emotion of this movement behavior, imitate this movement pattern or eventually adjust its own behavior. The aim of the overall project is to teach or to influence behavior of the human in order to improve his or her social skills.

The implementation of detection and tracking a waving hand is given in Figure 1. It consists of the following functional blocks that can be provided to the reader upon request:

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

The theoretical background and the computational details behind the implemented signal processing is described in more detail in the following subsections.

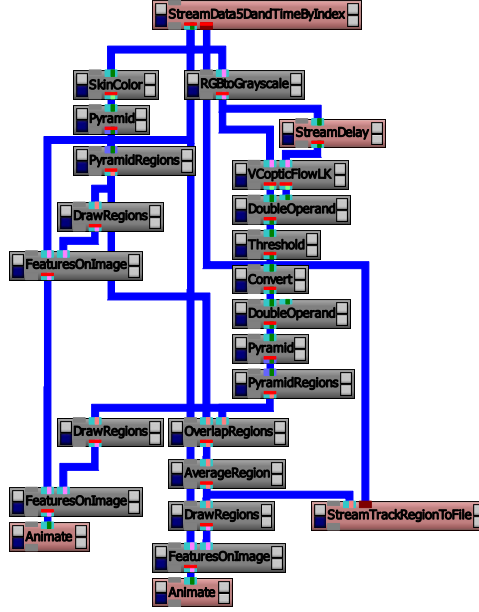


Figure 1. TiViPE (www.tivipe.com) implementation of handwaving. The icons from top to bottom at the left-side process skin areas, while motion sensitivity is processed by the functional blocks at the right-side.

2.1 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 $b \in [0, \dots, 255]$ 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

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

Threshold values as used by Chai and Ngan[6]

$$77 < Cb < 127, \quad 133 < Cr < 173$$

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:

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

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 1 yields a binary image.

The functional concept as described above contains similarities with how the brain processes visual data, especially in the way the primary visual cortex area V4 provides a substantial role in processing color [20].

2.2 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:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t). \quad (2)$$

In this method the following needs to be solved:

$$I_x V_x + I_y V_y = -I_t, \quad (3)$$

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 [17] is used, it is a two-frame differential method for optical flow estimation where the derivatives are obtained by using a Sobel filter kernel [18]. Instead of using Sobel kernels the more biologically plausible Gabor kernels can be used as well [7,11]. 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 [9,10]. It is plausible that a similar activity flow map might be expected in the middle temporal area MT, also known as primary visual cortex area V5 [20].

From (V_x, V_y) the L-2 norm (top right "DoubleOperand" icon of Figure 1) is taken and thresholded at 5 to obtain a binary "motion classified" image.

2.3 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 the pyramid is computed as follows:

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

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 $(I_i(2x, 2y), I_i(2x + 1, 2y), I_i(2x, 2y + 1), \text{ and } I_i(2x + 1, 2y + 1))$ 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:

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

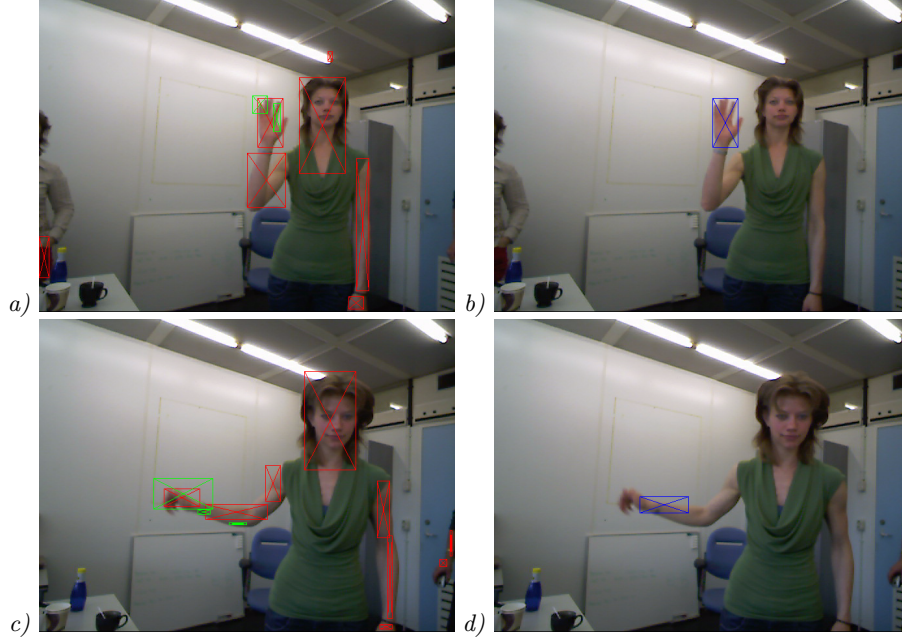


Figure 2. Marked regions of interest. Red and green areas denote skin and motion area, respectively. A blue area is the combined moving skin area that has been averaged over all skin areas that are moving.

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

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

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 2.

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

Formally such a feature f can be described by its region, type, and time: $f(x_l, x_r, y_u, y_d, \text{regiontype}, t)$. This f in turn could be further processed by other visual areas or passed on to both STS and PFC.

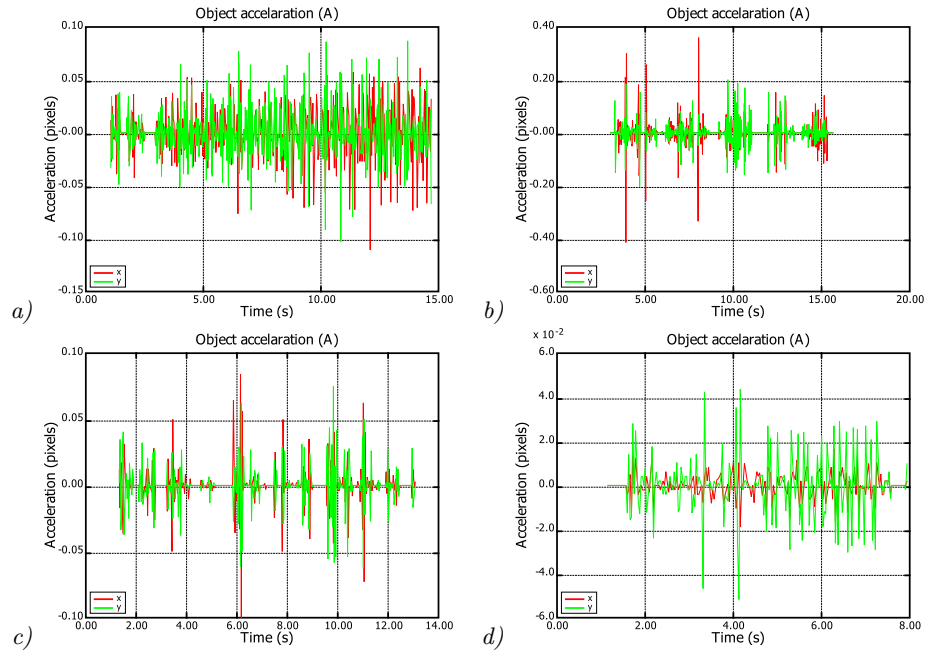


Figure 3. Waving patterns. First row shows acceleration profiles for happiness, and anger. The second row provides sadness, and politeness profiles.

2.4 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 2. Creating a single region in the image rather than multiple regions of interest is accomplished in order to unambiguously track an object of interest. These tracked objects are stored as file, see also icon “StreamTrackRegionToFile” of Figure 1, and processed further.

Fifteen recordings of 20 seconds have been made where a performer was asked to demonstrate happiness, anger, sadness, or politeness. In each plot the acceleration profile has been obtained by taking the second derivative of the central point of the tracked object. Examples for all four different emotional states are depicted in Figure 3. 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.

In an average acceleration-frequency plot of the recorded movements four distinctive clusters are formed (Figure 4). In one of the image sequences the actor 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 in Figure 3b shows an average acceleration score of more than 0.07 and gives an indication of the upper bound of used energy by performing these emotions by the actor.

3 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 2.4. In a complementary study we are attempting to classify motion by so-called Laban primitives [2]. Using these primitives we classify the intend and the mental state of the person that perform movement behavioral patterns.

The current method is developed to enable a robot to interact with a human in realistic scenarios. If a robot is able to track in parallel regions of interest, a considerable number of interacting scenarios are possible even without interpreting 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 a person that interacts with the robot performs repeatedly a movement with his/her head while waving, this will be learned and eventually included in the imitation behavior of the robot.

4 Discussion and Future work

In this paper we have shown that a device or robot is able to detect a waving hand in real time by using a combination of skin tone and motion. Tracking a hand motion provides clear insight about the emotional state of a person when displayed in frequency-amplitude domain. The four tested emotions are clearly clustered and segregated in this domain making the current approach suitable for emotion recognition. In a complementary study these regions are transferred 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 the overall research that 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?

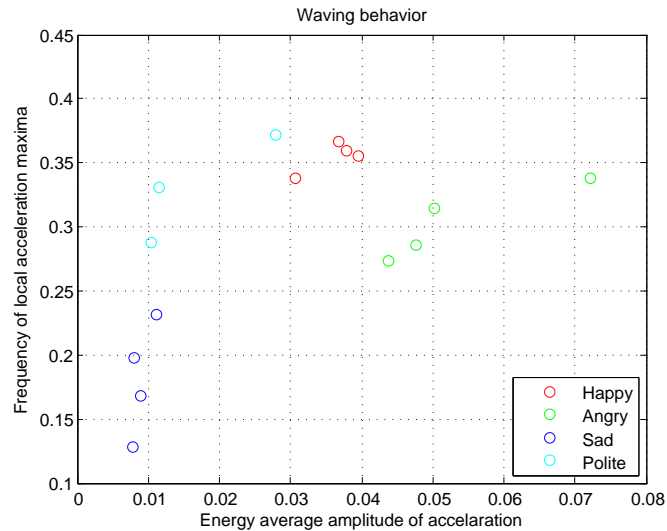


Figure 4. Distinct emotion profiles are revealed by average frequency and acceleration.

- How does the design of simple interaction behavior look like?
- How to design imitation, such that it appears natural and distinctive on a humanoid robot?

An important design aspect of a humanoid robot will be how closely human movement can be emulated on it. The emulation will be restricted by the understanding the physical limitations of the robots, and the mechanism that cause the movement behavior. Next, a basic set of motion primitives needs to be derived that independently of the physical embodiment emulates body movements that are interpreted by humans as emotional and social.

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. D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148:574–591, 1959.
10. 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.
11. T. Lourens. *A Biologically Plausible Model for Corner-based Object Recognition from Color Images*. Shaker Publishing B.V., Maastricht, The Netherlands, March 1998.
12. T. Lourens. Tivipe – tino's visual programming environment. In *The 28th Annual International Computer Software & Applications Conference, IEEE COMPSAC 2004*, pages 10–15, 2004.
13. 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.
14. T. Lourens and E. I. Barakova. Orientation contrast sensitive cells in primate v1 – a computational model. *Natural Computing*, 6(3):241–252, September 2007.
15. 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.
16. 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.
17. 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.
18. I. E. Sobel. *Camera Models and Machine Perception*. PhD thesis, Electrical Engineering Department, Stanford University, Stanford, CA, 1970.
19. 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.
20. S. Zeki. *A Vision of the Brain*. Blackwell science Ltd., London, 1993.