Human Machine Interaction based on Bayesian Analysis of Human Movements

Jörg Rett & Jorge Dias
Institute of Systems and Robotics, University of Coimbra, Polo II, 3030-290 Coimbra, Portugal

We present as a contribution to the field of human-machine interaction a system that analyzes human movements online, based on the concept of Laban Movement Analysis (LMA). The implementation uses a Bayesian model for learning and classification, while the results are presented for the application to gesture recognition. Nowadays technology offers an incredible number of applications to be used in human-machine interaction. Still, it is difficult to find implemented cognitive processes that benefit from those possibilities. Future approaches must offer to the user an effortless and intuitive way of interaction. We present the Laban Movement Analysis as a concept to identify useful features of human movements to classify human actions. The movements are extracted using both, vision and magnetic tracker. The descriptor opens possibilities towards expressiveness and emotional content. To solve the problem of classification we use the Bayesian framework as it offers an intuitive approach to learning and classification. It also provides the possibility to anticipate the performed action given the observed features. We present results of our system through its embodiment in the social robot 'Nicole' in the context of a person performing gestures and 'Nicole' reacting by means of audio output and robot movement.

1 Introduction

Nowadays, robotics has reached a technological level that provides a huge number of input and output modalities. Apart from industrial robots, also social robots have emerged from the universities to companies as products to be sold. The commercial success of social robots implies that the available technology can be both, reliable and cost efficient. Surprisingly, higher level cognitive systems that could benefit from the technological advances in the context of human-robot interaction are still rare. Future approaches must offer an effortless and intuitive way of interacting with a robot to its human counterpart. One can think of the problem as a scenario where a robot is observing the movement of a human and is acting according to the extracted information (see fig. 1). To achieve this interaction we need to extract the information contained in the observed movement and relate a appropriate robot action to it.

Our ultimate goal is to provide the robot with a cognitive system that mimics human perception in terms of anticipation and empathy. Towards the latter requirement this article will present the concept of Laban Movement Analysis (LMA) (Bartenieff and Lewis 1980) as a way to describe intentional content and expressiveness of a human body movement. Two major components of LMA (i.e. *Space* and *Effort*) are described in detail. We show the technical realization of LMA for the cognitive system of the embodied agent which is based on a probabilistic (Bayesian) framework and a system for tracking of human movements. The system uses both a magnetic tracker and a visual tracker. The visual tracker extracts the movement-features of a human actor from a series of images taken by a single camera. The hands and the face of the actor are detected and tracked automatically without using a special device (markers) (Rett and Dias 2006).

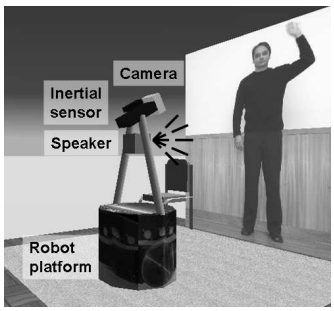

Figure 1: Nicole in position to interact.

This work presents the Bayesian approach to LMA through the problem of learning and classification, also treating the system's online characteristic of anticipation. The probabilistic model anticipates the gesture given the observed features using the Bayesian framework. The system has been implemented in our social robot, 'Nicole' to test several human-robot interaction scenarios (e.g. playing).

If the perceptual system of a robot is based on vision, interaction will involve *visual human motion analysis*. The ability to recognize humans and their activities by vision is key for a machine to interact intelligently and effortlessly with a human-inhabited environment (Gavrila 1999). Several surveys on visual analysis of human movement have already presented a general framework to tackle this problem (Aggarwal and Cai 1999), (Gavrila 1999), (Pentland 2000) and (Moeslund and Granum 2001) usually emphasizing the three main problems: 1. Feature Extraction, 2. Feature Correspondence and 3. High Level Processing. One area of high level analysis is that of gesture recognition applied to control some sort of devices. In (Pavlovic 1999) DBNs were used to recognize a set of eleven hand gestures to manipulate a virtual display shown on a projection screen . Surveys specialized on gesture interfaces along the last ten years reflect the development and achievements (Pavlovic, Sharma, and Huang 1997), (Moeslund and Norgard 2003). The most recent survey (Moeslund, Hilton, and Kruger 2006) is once more included in the broader context of human motion analysis emphasizing, once more the dependencies between low level features and high level analysis.

Section 2 presents the concept of LMA and its two major components (i.e. *Space*and *Effort*). Section 3 presents the system for tracking of human movements. Section 4 describes the Bayesian framework that is used to learn and classify human movements and presents the. Section 5 presents the results. Section 6 closes with a discussion and an outlook for future works.

2 Laban Movement Analysis (LMA)

Laban Movement Analysis (LMA) is a method for observing, describing, notating, and interpreting human movement. It was developed by a German named Rudolf Laban (1879 1958), who is widely regarded as a pioneer of European modern dance and theorist of movement education (Zhao 2002). While being widely applied to studies of dance and application to physical and mental therapy (Bartenieff and Lewis 1980), it has found little application in the engineering domain. Most notably the group of Norman Badler, who recently proposed a computational model of gesture acquisition and synthesis to learn motion qualities from live performance (Zhao and Badler 2005). Also recently, researchers from neuroscience stated that LMA is quite useful to describe certain effects on the movements of animals and humans. In (Foroud and Whishaw 2006) LMA was adapted to capture the kinematic and non-kinematic aspects of movement in a reach-for-food task by human patients whose movements had been affected by stroke.

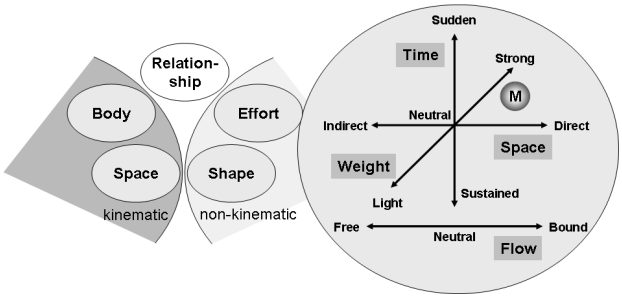

Figure 2: Major components of LMA with the bipolar Effort factors as a 4-D space

The theory of LMA treats five major components shown in fig. 2 of which we adopted three. *Space* treats the spatial extent of the mover's *Kinesphere* (often interpreted as reach-space) and what form is being revealed by the spatial pathways of the movement. *Effort* deals with the dynamic qualities of the movement and the inner attitude towards using energy. Like suggested in (Foroud and Whishaw 2006) we have grouped *Body* and *Space* as kinematic features describing changes in the spatial-temporal body relations, while *Shape* and *Effort* are part of the non-kinematic features contributing to the qualitative aspects of the movement.

2.1 Space

The *Space* component addresses what form is being revealed by the spatial pathways of the movement. The actor is actually "carving shapes in space" (Bartenieff and Lewis 1980). *Space* specifies different entities to express movements in a frame of reference determined by the body of the actor. Thus, all of the presented measures are relative to the anthropometry of the actor. The concepts differ in the complexity of expressiveness and dimensionality but are all of them reproducible in the 3-D Cartesian system. The most important ones shown in fig. 3 are: 1) The *Levels of Space* - referring to the height of a position, 2) The *Basic Directions* - 26 target points where the movement is aiming at, 3) The *Three Axes* - Vertical, horizontal and sagittal axis, 4) The *Three Planes* - *Door Plane* π_v, *Table plane* π_h, and the *Wheel Plane* π_s each one lying in

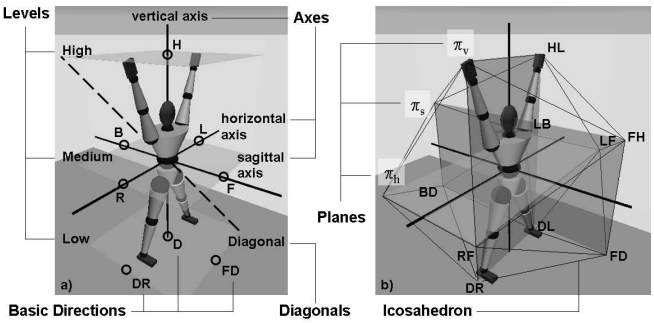

Figure 3: The concepts of a) Levels of Space, Basic Directions, Three Axes, and b) Three Planes and Icosahedron

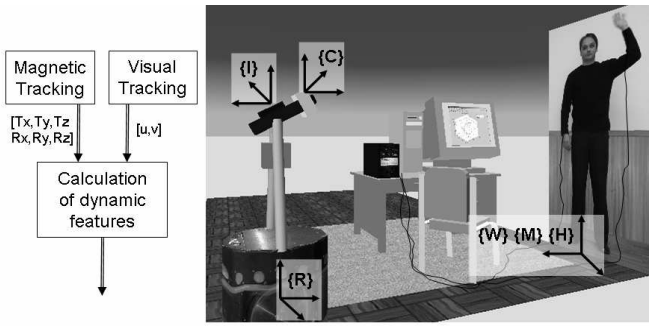

Figure 4: The components and the frames of reference for tracking human movements.

two of the axes, and 5) The *Icosahedron* - used as *Kinespheric Scaffolding*. The *Kinesphere* describes the space of farthest reaches in which the movements take place. Levels and Directions can also be found as symbols in modern-day Labanotation (Bartenieff and Lewis 1980). Labanotation direction symbols encode a position-based concept of space. Recently, Longstaff (Longstaff 2001) has translated an earlier concept of Laban which is based on lines of motion rather than points in space into modern-day Labanotation. Longstaff coined the expression *Vector Symbols* to emphasize that they are not attached to a certain point in space. The 38 *Vector Symbols* are organized according to *Prototypes* and *Deflections*. The 14 *Prototypes* divide the Cartesian coordinate system into movements along only one dimension (*Pure Dimensional Movements*) and movements along lines that are equally stressed in all three dimensions (*Pure Diagonal Movements*) as shown in fig. 3 a). Longstaff suggests that the *Prototypes* give idealized concepts for labeling and remembering spatial orientations. The *Vector Symbols* are reminiscent of a popular concept from neuroscience, named *preferred directions*, which are the directions that trigger the strongest response from motion encoding cells in visual area MT of a monkey (Pouget, Dayan, and Zemel 2000).

2.2 Effort

The *Effort* component consists of four motion factors: *Space*, *Weight*, *Time*, and *Flow*. As each factor is bipolar and can have values between two extremities one can think of the *Effort* component as a 4-D space as shown in fig. 2. A movement (M) can be described by its location in the *Effort*-space. Exemplary movements where a certain *Effort*-value is predominant are given in table 1. It is important to remember, that a movement blends during each phase all four Effort-value. Most of the human movements have two or three *Effort*-values prominently high. In fact it, seems difficult even for a trained Laban performer (i.e. Laban notator) to perform single-quality movements (Zhao 2002).

3 Tracking of human movements

For the tracking of human movements we use sensory data from a camera, which is mounted on our social robot, Nicole and a magnetic tracker as shown in fig. 4. From the camera we collect 2-D position data of the hands and head with 15Hz. The magnetic tracker produces 3-D position and orientation data with 50Hz for each sensor. The number of sensors and their location depends on the performed action (e.g three sensors on hands and head for gestures). We have created a database of human movements, called HID-Human Interaction Database which is publicly accessible through the internet (Rett, Boussier, Sousa, Neves, Faria, and Dias 2007). HID is organized in three categories of movements: 1. Gestures (e.g waving bye-bye), 2. Expressive movements in terms of LMA as presented in tab. 1 (e.g. performing a punch) and 3. Manipulatory movements performing reaching, grasping and placing of objects (e.g. drinking from a cup). Figure 4 indicates some of the frames of references involved: The camera referential C in which the image is defined, the inertial referential I allowing us to register the image data in the vertical and the robot referential R which defines the position and orientation of the visual system relative to some world coordinate system W. In the current situation the frame of reference of the

Effort	Movement
Space Direct	Pointing gesture
- Indirect	Waving away bugs
Weight Strong	Punching,
- Light	Dabbing paint on a canvas
Time Sudden	Swatting a fly
- Sustained	Stretching to yawn
Flow Bound	Moving in slow motion
- Free	Waving wildly

Table 1: *Effort* qualities and exemplary movements

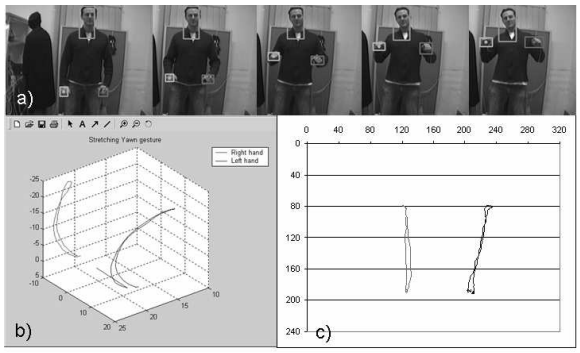

Figure 5: Tracking of hands movement. a) Sample images b) Data from the magnetic tracker c) ... and the vision tracker

world W coincides with the on of the magnetic tracker M and the one we contribute to the human H.

3.1 Tracking using 6-DoF magnetic tracker

Using a 6-DoF magnetic tracker provides 3-D position data with a sufficiently high accuracy and speed. We use a Polhemus Liberty system with sensors attached to several body parts and objects. From the tracker data set of features is calculated and related to the Laban Movement Parameters (LMP). Figure 5 a) shows some sample images from the expressive movement "Stretching to yawn" and in fig. 5 b) the trajectories for both hands. The tracker data is used to learn the dependencies of the features from the LMPs. Subsets (e.g. 2-D vertical plane) are used to test the expressiveness in lower dimensionality like vision.

3.2 Tracking using vision

Using cameras as the basic input modality for a robot provides the highest degree of freedom to the human actor but also poses the biggest challenge to the functionality of detecting and tracking of human movements. To collect the data we use the gesture perception system (GP-System) (Rett and Dias 2005) of our social robot Nicole. The system performs skin-color detection and object tracking based on the CAMshift algorithm presented in (Bradski 1998). From the position data the displacement vectors dP between each frame are calculated. The spatial concept of Laban's *Vector Symbols* is implemented by defining a finite number of discrete values for the direction and calculating what we call *Vector Atoms* or simply *Atoms A*.

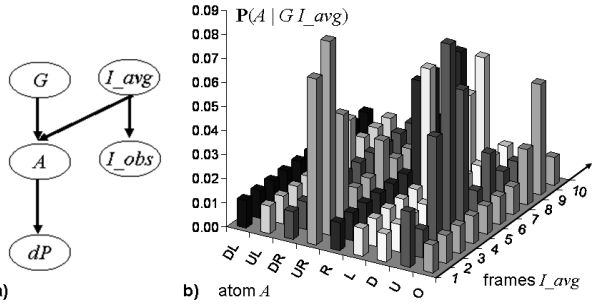

Figure 6: a) Bayesian Net for the gesture model b) Learned Table $P(A|GI\_avg)$ for gesture 'Bye-Bye'.

4 Bayesian Framework for Movement Classification

The classification of human movements is done with a probabilistic model using a Bayesian framework. The Bayesian framework offers a broad range of possible models (HMMs etc.) and has proven successful in building computational theories for perception and sensorimotor control of the human brain (Knill and Pouget 2004). These models have already shown their usability in gesture recognition (Starner 1995; Pavlovic 1999).

The model for Laban *Space* uses as input (evidences) the *Atoms A*. Our solution assumes that the probability distribution for all possible values of atom A given all possible gestures G and frames I ,which is $\mathbf{P}(A|G, I)$ can be determined. As both, the gestures and the frame index are discrete values we can express $P(A|G, I)$ in form of a conditional probability table. The probabilities can be learned from training data using a certain number of atom-sequences for each gesture. A simple approach is the one known as Histogram-learning. It counts the number of different atom-values that appear for a gestures along the frames. To overcome the problem of assigning zero probabilities to events that have not yet been observed an enhanced version often uses learning of a family of Laplace-distributions. Currently we are using a table that is of size 18 x 31 x 6, that is 18 discrete values for the atom (9 for each hand), 31 frames and 6 gestures. Figure 6 shows a fraction of the table which is the 9 atoms of the right hand for the first 11 frames and the Bye-Bye gesture.

It represents the 'fingerprint' of the gesture prototype for waving Bye-Bye. Knowing the gesture we assume this sequence of distributions of the random variable atom to extracted. The table represents an intuitive way to distinguish two gestures from each other.

Applying Bayes rule we can compute the probability distribution for the gestures G given the frame I and the atom A expressed as $\mathbf{P}(G|I, A)$, which is the question the classification as based

upon. $\mathbf{P}(G)$ represents the prior probabilities for the gestures. Assuming the the observed atoms are independently and identically distributed (i.i.d.) we can compute the probability that a certain gesture has caused the whole sequence of atoms $P(a_{1:n}|g, i_{1:n})$ by the product of the probabilities for each frame. Where $a_{1:n}$ represents the sequence of n observed values for atom and g a certain gesture from all gestures G. The jth frame of a sequence of n frames is represented by i_j. We are able to express the probability of a gesture g that might have caused the observed sequence of atoms $a_{1:n}$ in a recursive way. Assuming that each frame a new observed atom arrives we can state and expressing the real-time behavior by using the index t. We model the variance and mean speed of a performance by a Gaussian distribution $N(i\_obs, \sigma)$ expressed the probability that an observed frame i_obs maps to an average frame $iavgP(i_obs|i_avg)$.

Our Bayesian model is shown in equation 1. We see that the probability distribution of the gestures G at time $t+1$ knowing the observed atoms a until $t+1$ is equal to the probability distribution of G at time t times the probabilities of the current observed atom given the gestures G and frame i at $t+1$. The probability distribution of G for $t=0$ is the prior.

$$\mathbf{P}(G_{t+1}|i_{1:t+1}, a_{1:t+1}) = \mathbf{P}(G_t)P(i\_obs_{t+1}|i\_avg_{t+1})\mathbf{P}(a_{t+1}|G, i_{t+1}) \quad (1)$$

We can likewise express our model in a *Bayesian Net* shown in fig. 6. It shows the dependencies of the above mentioned variables including the displacement dP from the previous section. The rule for classification is based on the highest probability value above a minimum threshold, also known as maximum a posteriori estimation (MAP).

5 Results and Discussion

For this experiment we have used 15 video sequences from each human actor for each of 6 distinct gestures as shown in table 2. Figure 7 illustrates how the gesture-hypothesizes, evolve as new evidences (atoms) arrive taken from the performance of a Bye-Bye gesture. After twelve frames the probabilities have converged to the correct gesture-hypothesis (No. 4). After four frames the probabilities of the two-hand gesture-hypothesis have reached nearly zero. (No. 1, 2, and 3). Until the sixth frame the probabilities of both *High-Level* gestures grow (No. 4 and 5) indicating what is called pre-stroke phase in gesture analysis (Rossini 2003). Conversely the probability of the *Medium-Level* gesture (No. 6) drops slowly towards zero. After the sixth frame the oscillating left-right movement (and its associated atoms) makes the probability of the Bye-Bye-gesture hypothesis rise and the Pointing-NW-gesture hypothesis drop. A similar behavior was revealed when the remaining five gestures were performed. An unknown gesture, i.e. an unknown sequence of atoms produced more than one gesture-hypothesizes with a significant probability.

No.	Gesture	Hands	Level
1	Sagittal Waving	Two	High
2	Waving to Left	Two	Medium
3	Waving to Right	Two	Medium
4	Waving Bye-Bye	One	High
5	Pointing	One	High
6	Draw Circle	One	Medium

Table 2: Characteristics of out gesture-set

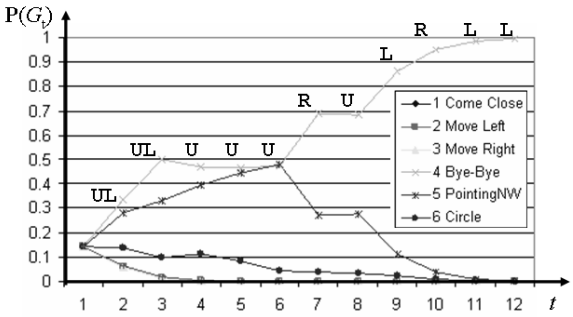

Figure 7: Probability evolution for a Bye-Bye gesture input.

For the Bye-Bye gesture (see fig. 6) we can see, that during the first frames the most likely atom to be expected is the one that goes Up-Right (UR). This is similar for the Pointing gesture (see fig. 8) reflecting the already mentioned *Pre-Stroke* phase. The number of atoms during *Pre-Stroke* also reflect the *Levels of Space* in which the following *Stroke* (Rossini 2003) will take place. In our example we can distinguish the two gestures during

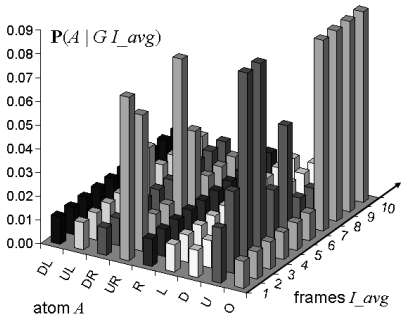

Figure 8: Learned Table $P(A|GI\_avg)$ for gesture 'Pointing NW'.

Stroke as the Bye-Bye gesture has a roughly equal distribution along the line of oscillation (e.g. left-right), while the Pointing gesture produces mainly zero-motion atoms (O).

6 Conclusions and Future Works

This work presented the application of the *Space* component of Laban Movement Analysis (LMA) to the Human-Robot Interface of the social robot, Nicole. It showed that trajectories of human movements can be learned and recognized using the concept of *Vector Symbols*. This work demonstrates that the *Bayesian approach for movement classification* provides a robust and reliable way to classify gestures in real-time. Using naive Bayesian classification we are able to anticipate a gesture from its beginning and can take decisions long before the performance has ended. We have shown that through *Bayesian Learning* the system memorizes learned data in an intuitive way which gives the possibility to draw conclusions directly from the look-up tables. In several trials the system was successfully performing Human-Robot Interaction with guests and visitors.

The next step will be the application of the *Effort* and *Shape* component of the LMA to Nicole. Incorporating the dynamic qualities we hope to classify also the emotional expression of a human movement. For the future we aim to continue to develop a social platform where the impact of imitation between a human and a machine can be observed.

7 Acknowledgements

This work is partially supported by FCT-Fundação para a Ciência e a Tecnologia Grant #12956/2003 to J. Rett and by the BACS-project-6th Framework Programme of the European Commission contract number: FP6-IST-027140, Action line: Cognitive Systems

REFERENCES

Aggarwal, J. K. and Q. Cai (1999). Human motion analysis: A review. *CVIU 73*(3), 428–440.

Bartenieff, I. and D. Lewis (1980). *Body Movement: Coping with the Environment*. New York: Gordon and Breach Science.

Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal 2*(Q2), 15.

Foroud, A. and I. Q. Whishaw (2006). Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke: A laban movement analysis in two case studies. *Journal of Neuroscience Methods 158*, 137–149.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *CVIU 73*(1), pp. 82–98.

Knill, D. C. and A. Pouget (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences 27*, 712–719.

Longstaff, J. S. (2001). Translating "vector symbols" from labans (1926) choreographie. In *26. Biennial Conference of the International Council of Kinetography Laban, ICKL, Ohio, USA*, pp. 70–86.

Moeslund, T., A. Hilton, and V. Kruger (2006, November). A survey of advances in vision-based human motion capture and analysis. *CVIU 103*(2-3), 90–126.

Moeslund, T. B. and E. Granum (2001). A survey of computer vision-based human motion capture. *CVIU 81*(3), 231–268.

Moeslund, T. B. and L. Norgard (2003). A brief overview of hand gestures used in wearable human computer interfaces. Technical report, Computer Vision and Media Technology Lab., Aalborg University, DK.

Pavlovic, V., R. Sharma, and T. S. Huang (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*(7), 677–695.

Pavlovic, V. I. (1999). *Dynamic Bayesian Networks for Information Fusion with Applications to Human-Computer Interfaces*. Ph. D. thesis, Graduate College of the University of Illinois.

Pentland, A. (2000, January). Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on PAMI 22*(1), 107–119.

Pouget, A., R. Dayan, and R. Zemel (2000). Information processing with population codes. *Nature Reviews Neuroscience 1*, 125132.

Rett, J., E. Boussier, B. Sousa, A. Neves, D. Faria, and J. Dias (2007). Hid-human interaction database: http://paloma.isr.uc.pt/pub/bscw.cgi/0/110272.

Rett, J. and J. Dias (2005). Visual based human motion analysis: Mapping gestures using a puppet model. In *Proceedings of EPIA 05, Lecture Notes in AI, Springer Verlag, Berlin.*

Rett, J. and J. Dias (2006). Gesture recognition using a marionette model and dynamic bayesian networks (dbns). In *Proceedings of ICIAR 2006, Lecture Notes in CS, Springer Verlag, Berlin*, Volume 4142/2006, pp. 69–80.

Rossini, N. (2003). The analysis of gesture: Establishing a set of parameters. In *Gesture-based Communication in Human-Computer Interaction, LNAI 2915, Springer Verlag*, pp. 124–131.

Starner, T. (1995, Feb). Visual recognition of american sign language using hidden markov models. Master's thesis, MIT.

Zhao, L. (2002). *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures*. Ph. D. thesis, University of Pennsylvania.

Zhao, L. and N. I. Badler (2005, January). Acquiring and validating motion qualities from live limb gestures. *Graphical Models 67*(1), 1–16.