Cloning with gesture expressivity

Manoj Kumar Rajagopal

Thèse de doctorat de Télécom SudParis
dans le cadre de l'école doctorale S&I en co-accréditation avec

l'Université d'Évry-Val d'Essonne

Spécialité :

Informatique

Par

Manoj Kumar Rajagopal

Thèse présentée pour l'obtention du diplôme de Docteur

de Télécom SudParis

Clonage gestuel expressif

Soutenue le 11 mai 2012 devant le jury composé de

M. Patrice Dalle, Professeur à l'Université Paul Sabatier de Toulouse, rapporteur
M. Rachid Deriche, Directeur de Recherche,
INRIA Sophia Antipolis-Méditerranée, rapporteur
Mme Sylvie Lelandais, Professeur à l'Université d'Évry-Val d'Essonne, examinatrice
Mme Bernadette Dorizzi, Professeur à Télécom SudParis, directrice de thèse
Mme Catherine Pelachaud, Directrice de Recherche,
CNRS - LTCI Télécom ParisTech, co-directrice de thèse
M. Patrick Horain, Ingénieur d'Études à Télécom SudParis, encadrant

Thèse n°2012TELE0013

PhD Thesis prepared at Telecom SudParis
in the framework of École doctorale S&I in partnership with

University of Evry-Val d'Essonne

Specialized in:

Computer Science

By

Manoj Kumar Rajagopal

A dissertation submitted for the degree of Doctor of Philosophy

at Telecom SudParis

Cloning with gesture expressivity

Defended on 11th May 2012 before the jury composed of:

M. Patrice Dalle, Professor at University Paul Sabatier in Toulouse, reviewer
M. Rachid Deriche, Director of Research,
INRIA Sophia Antipolis-Méditerranée, reviewer
Mme Sylvie Lelandais, Professor at University of Evry-Val d'Essonne, examiner
Mme Bernadette Dorizzi, Professor at Telecom SudParis, thesis director
Mme Catherine Pelachaud, Director of Research,
CNRS - LTCI Télécom ParisTech, co-directrice de thèse
M. Patrick Horain, Associate Professor at Telecom SudParis, advisor

Thèse no: 2012TELE0013

Abstract

Virtual environments allow human beings to be represented by virtual humans or avatars. Users can share a sense of virtual presence is the avatar looks like the real human it represents. This classically involves turning the avatar into a clone with the real human's appearance and voice. However, the possibility of cloning the gesture expressivity of a real person has received little attention so far. Gesture expressivity combines the style and mood of a person. Expressivity parameters have been defined in earlier works for animating embodied conversational agents.

In this work, we focus on expressivity in wrist motion. First, we propose algorithms to estimate three expressivity parameters from captured wrist 3D trajectories: repetition, spatial extent and temporal extent. Then, we conducted perceptual study through a user survey the relevance of expressivity for recognizing individual human. We have animated a virtual agent using the expressivity estimated from individual humans, and users have been asked whether they can recognize the individual human behind each animation.

We found that, in case gestures are repeated in the animation, this is perceived by users as a discriminative feature to recognize humans, while the absence of repetition would be matched with any human, regardless whether they repeat gesture or not. More importantly, we found that 75 % or more of users could recognize the real human (out of two proposed) from an animated virtual avatar based only on the spatial and temporal extents. Consequently, gesture expressivity is a relevant clue for cloning. It can be used as another element in the development of a virtual clone that represents a person.

Résumé

Les environnements virtuels permettent de représenter des personnes par des humains virtuels ou avatars. Le sentiment de présence virtuelle entre utilisateurs est renforcé lorsque l'avatar ressemble à la personne qu'il représente. L'avatar est alors classiquement un clone de l'utilisateur qui reproduit son apparence et sa voix. Toutefois, la possibilité de cloner l'expressivité des gestes d'une personne a reçu peu d'attention jusqu'ici. Expressivité gestuelle combine le style et l'humeur d'une personne. Des paramètres décrivant l'expressivité ont été proposés dans des travaux antérieurs pour animer les agents conversationnels.

Dans ce travail, nous nous intéressons à l'expressivité des mouvements du poignet. Tout d'abord, nous proposons des algorithmes pour estimer trois paramètres d'expressivité à partir des trajectoires dans l'espace du poignet : la répétition, l'étendue spatiale et l'étendue temporelle. Puis, nous avons mené une étude perceptive sur la pertinence de l'expressivité des gestes pour reconnaître des personnes. Nous avons animé un agent virtuel en utilisant l'expressivité estimée de personnes réelles, et évalué si des utilisateurs peuvent reconnaître ces personnes à partir des animations.

Nous avons constaté que des gestes répétitifs dans l'animation constituent une caractéristique discriminante pour reconnaître les personnes, tandis que l'absence de répétition est associée à des personnes qui répètent des gestes ou non. Plus important, nous avons trouvé que 75% ou plus des utilisateurs peuvent reconnaître une personne (parmi deux proposée) à partir d'animations virtuelles qui ne diffèrent que par leurs étendues spatiales et temporelles. L'expressivité gestuelle apparaît donc comme un nouvel indice pertinent pour le clonage d'une personne.

Acknowledgments

Since this PhD adventure started, more than 3 years ago, I have been very fortunate to meet and work with many great people. Here, I would like to express my sincere gratitude to all of them who made the development of this thesis possible.

First and foremost I would like to thank my supervisor Patrick Horain, for choosing me to work on this interesting thesis. I am greatly indebted to his perfect mentoring, and support – both technical and personal. Interactions with him are always positive and fruitful. I would always cherish this 3½ years experience and take back plenty of wisdom that he has happily shared with me.

Another important person whom I would like to thank is my thesis Co-director, Catherine Pelachaud. During starting stage of thesis, she gave me variety of research publications to start my work. She showed me right direction whenever I struggled to proceed further in my thesis. I would also thank to her team members Andre-Marie Pez and Radoslaw Niewiadomski for giving me the technical support in using Greta and providing valuable inputs in my work.

Thanks to the jury members of my thesis for their valuable discussion during the thesis defense. Special thanks to Patrice Dalle and Rachid Deriche for their precise reviews and contributions which allowed enriching this thesis work.

I would like also to thank my thesis director, Bernadette Dorizzi, for accepting me in the EPH department of Télécom SudParis, where I found a very pleasant and great working environment during these years. Special thanks to Patricia Fixot and Marie-Thérèse Courcier, for her friendship and patience with me despite my several distractions. Thanks for helping me efficiently with the administrative problems that I had during these years.

I also would like to thank my wife Priya. Her constant encouragement and sacrifices need special mention here. I attribute a lot of my character to my parents. Thanks to them for that.

I need to give a special thanks to David Gomez, who is my senior in my lab. He helped me a lot when I faced language problem. He is my unofficial translator. Thanks a lot to David.

I am very grateful with all my laboratory colleagues and friends (Daniel, , Sesh, Maher, Daria, Quocdinh, Jerome) for the great time spent, inside and outside the lab, and their valuable help and support during the past years. I am also very grateful with all the good friends in the maisel and out of the maisel that I met during my PhD and whom I spent great times making my stay in France a wonderful experience.

Doing PhD has its ups and downs. I express my sincere apologies if I had hurt someone’s feelings or expectations during the last four years. Hope my thesis is at least a drop in the ocean of knowledge created by numerous research scholars over hundreds of years.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1 Problem Statement

Online virtual worlds allow multiple users to interact remotely by means of animated virtual representations, including virtual humans or avatars. They allow for interaction and collaboration around 3D objects.

Virtual humans are a new kind of intelligent Human-computer interface that allow human like animation and conversational skills (Cassell, Sullivan, Prevost, & Churchill, 2000). They may exhibit a human like aspect, both in appearance and behavior, including emotional states, personality traits etc. Virtual humans imply many complex problems that have been studied for years (Thalmann & Thalmann, 1991). Animating virtual humans with actions that reflect real human motion involves challenges (Thalman, 2000) such as controlling limbs deformations and collisions, high-level direction of avatars, adaptation of pre-defined movements and interacting with objects. Badler *et al*. (1999) state that animating a virtual human involves generating movements, reactions and interactions that appear "natural", appropriate and contextually sensitive.

In order to interact socially with a real human, a virtual human needs to behave like a real human. A virtual human should present the appearance and voice of a real human. It should also gesture like a real human. The process of making a virtual human look like a real human is termed as cloning.

The behavior of the person depends on personal general tendencies, or "style" (Gallaher, 1992). Most of the approaches to style are based on low level parameters like joint angles (Tenanbaum & Freeman, 2000) (Elgammal & Lee, 2004) (Grochow, Martin, Hertzmann, & Popović, 2004), but, practically it is difficult to address the style in terms of joint angles or any other low level parameters. Cloning should also involve gesturing style.

Non-verbal behavior conveys user's emotional states (Hassin, James, & Bargh, 2005). Mehrabian *et al.* (1967) says that body language expresses 55 % of the human feelings and intentions. Boone *et al.* (1998) work says hand gesture expresses emotional state of a person. Non-verbal behavior, body language and hand gestures play a major role in delivering human feelings. Human gestures are affected by their mood (Zhu & Thagard, 2002). The outcomes of the phase of action generation are mental representations such as intentions, decisions, choices, or goals. These representations are usually on the higher level of the hierarchical organization of action and therefore are relatively abstract (Gallistel, 1980).

This motivates to have a virtual human that reproduces the non-verbal behavior of a real human. We aim at capturing the style and mood of a person, which we jointly refer to as "expressivity" (Hartmann, Mancini, & Pelachaud, 2005), and render this with virtual human in such a way that it can be perceived and recognized by users.

1.2 Thesis contributions

In our work, we first propose algorithms to estimate three expressivity parameters from 3D human motion.

Then, through a user survey, we evaluate whether expressivity works as a clue for virtual cloning of humans. Using the Hartman *et al*. (2005) animation engine and conversational agent, we animated a virtual human with the expressivity estimated from real humans. The survey showed that up to 88 % of users could recognize the real human (out of two proposed) behind the virtual animation, based on expressivity only.

1.3 Organization of the Thesis

This thesis is organized as follows:

— Chapter 2 presents the state-of-the art on describing the style of human gestures, mood and expressive animation. It introduces parameters to describe expressivity in communicative gestures.

— In chapter 3, we propose algorithms to estimate three expressivity parameters from 3D motion trajectories.

— In chapter 4, experimentally evaluate whether users can recognize humans from virtual animations that encompass their estimated expressivity.

— Chapter 5 concludes with a summary of contributions and discusses future directions.

— The appendix shows the animation engines we employed in our animation process.

Chapter 2 - State of the Art

2.1 Introduction

Expressivity is conveyed by non verbal behavior of the person. . It conveys user's mental and affective states. In the process of defining expressivity we first introduce communicative gesture and then we discuss human expressivity. We address human expressivity and body language from the psychologist's point of view. We discuss the automated synthesis and analysis of human gestures from computer scientist's view. We also discuss the expressivity parameters and available animation engine to animate a virtual human.

2.2 Communicative Gesture – Definition

Communication takes place not only through speech, voice but also by means of gestures such as facial expressions, gaze, head movements, hand movements etc. As per Oxford dictionaries, the gesture is a movement of part of the body, especially a hand or the head, to express an idea or meaning. According to Poggi and Pelachaud, (2008) definition a communicative gesture stems from a notion of communication, based on a model in terms of goals and beliefs. Poggi (2007) defines communication as the case in which sender produces a perceivable signal by performing an action (a word, gesture, a glace of view etc.) or exhibiting a morphological trait in order to transform the intention to someone.

Based on this, Poggi and Pelachaud (2008) define communicative gesture as "a particular movement of hands, arms or shoulders that is used by a sender for the goal of communicating some meaning to some addressee" involving shape and positions. The meaning is a belief or a set of beliefs. A communicative gesture can be presented as a signal which aims at transmitting some meaning. In the process of analyzing individual person communicative gestures, we process the video of a person who is communicating with others. After analysis we define the human expressivity based on the hand motions during communication.

2.3 Gesture Expressivity

In face-to-face communication, body language can convey up to 55 % of the information on feelings and intentions (Mehrabian & Wiener, 1967). According to Mehrabian, among channels can convey feelings: words account for 7%, tone of voice accounts for 38%, and body language accounts for 55%. They are often abbreviated as the "3 Vs" for Verbal, Vocal & Visual. Facial expressions are a great mean to express emotions (Ekman, 1982). He states that some facial expressions of emotion are not culturally determined, but universal across human cultures and thus biological in origin. Some studies (Boone & Cunningham, 1998) (Meijer, 1989) decode the emotional states from expressive body movements. These studies reveal that how body expressions play a role in reflecting the expressivity of a person.

Expressivity is also present how body gesture is performed. One of the most influential systems for describing the body gesture and its transcription came from sign language studies. During 1960's William Stokoe (1960) first analyzed American Sign Language (ASL) sign in terms of a small set of parameters namely hand shape, hand movement and hand location. Later other parameters like, hand orientation, wrist orientation and arm position were added (Klima & Bellugi, 1979) (Prillwitz, Leven, Zienert, Hanke, & Henning, 1989). This way of analyzing sign language is also used to analyze gestures. Later temporal gestures came to limelight by other researchers (McNeill, 1992) (Kita, Gijn, & Hulst, 1998) (Kendon, 2004).

A gesture has an excursion, from when the hand leaves its resting position and up to its coming back to it. According to Efron's (1941) gestural theory, all gestures are classified in three phases videlicet preparation, stroke and retraction. The stroke is always necessarily present, and it is the phase of the excursion in which the shape of the gesture and the movement dynamics are clear (Kendon, 2004) Several researchers (Wallbott & Scherer, 1986), (Gallaher, 1992), (Ball & Breese, 2000), (Pollick, 2003) have investigated human motion characteristics and encoded them in to dimensional categories. Some authors refer to body motion using dual categories such as slow *vs*. fast, small *vs*. expansive, weak *vs*. energetic, small *vs*. large, pleasant *vs*. unpleasant. Bevacqua *et al.* (2007) define expressivity of behavior as the "How" the information is communicated through the execution of some physical behavior.

2.3.1 Expressivity – Psychologists Views

Dancers show their expression in hand motions while dancing. Their hand motions have lot of variations they dance. They also used to communicate some message while dancing. Laban (1960) created a way and language for interpreting, describing, visualizing and notating all ways of human movement called Laban Movement Analysis (LMA). LMA is composed of five major components namely body, space, effort, shape and relationship. Body deals with which body parts move, where the movements initiate and how the movement spreads through the body. Space describes how large the mover's kinesphere[1] and what crystalline form is being revealed by spatial pathways of the movement. Shape describes the changing forms that the body makes in space while effort involves the dynamic qualities of the movement and the inner attitude towards using energy. Relationship describes modes of interaction with oneself, others, and the environment. Each individual has their own unique repertoire and preferences for combinations of these basic elements which can be sequences, phrased, patterned and orderly organized together in a particular personal and artistic way.

Modern Psychological and a virtual human animation researches on hand gestures are based on systematic observations or experiments conducted by Efron (1941). According to Efron's gestural theory, in the preparation phase, the hands are raised to the location where the gesture begins. In the stroke phase, the actual gesture is performed and the hands relax and fall back to the resting places in the restoration phase.

Johansson (1973) worked with Moving Light Displays (MLD) affiliated with body parts showed that observers can almost identified biological motion patterns right away even when presented with only few of these moving dots. MLDs generate robust shape from motion cues that allow identity, gender, emotions and personality traits. This raised the question whether recognition of moving parts can be achieved directly from motion, without structure recovery.

Wallbott and Scherer (1986) classified the bodily behavior using five criterions: slow or fast, small or expensive, weak or energetic, small or large movement activity and unpleasant or pleasant. Behavior expressivity has been correlated to energy in communication, to the relation with characteristics of gestures and/or personality/emotion. Wallbott (1998) asked actors to portray fourteen different emotions for a given scenario. He aimed at characterizing emotions based on specific body movement and posture. A coding schema for body movement and posture was designed. Movement for each anatomy body part (hand, arm, head, shoulder, and upper body) is encoded as well as movement quality. For upper body, three dimensions are annotated namely 'movement activity' (overall quantity of movements), 'expansiveness/spatial extension' (of body parts), and 'movement dynamics/energy/power' (of body parts). Totally there are twenty six categories in

[1] Kinesphere: the area that the body is moving within and how the mover is paying attention to it.

the annotation schema. From the analysis of the annotation it was apparent that most of the categories (seventeen in total) served to differentiate emotions. In particular, the three movement quality dimensions showed significant differences for the fourteen emotions. For example, the emotion ‘boredom’ is shown by low movement activity, un-expansive movements, and low movement dynamics, while ‘elated joy’ is characterized by high movement activity, expansive movement, and high movement dynamics. This shows human movements encode not only content information (information communicated through gesture shape) but also emotional information (the how it is communicated, the manner of execution).

McNeil (1992) explains gestures plays a major role between our conceptualizing capacities and our linguistic abilities. Humans use a very wide variety of gestures ranging from simple actions of using the hand to point at objects to the more complex actions that express feelings and allow communication with others.

Allport and Vernan (1933) consider three main features to describe gesture expressivity, namely an aerial factor of broad versus constricted movements, a centrifugal factor of movements away from the body versus towards the body and an emphasis factor of forceful versus weak movement. Inter-individuality differences are found between subjects performing different task. But there is also a lot of intra individual consistency: individuals showed the same behavior quality over the different tasks they performed.

Gallaher (1992) found consistencies in the way people behave. She conducted evaluation studies in which subject’s behavior style was evaluated by friends, and by self-evaluation. In a first study (Gallaher, 1992), she evaluated many characteristics of behavior: tendency to use body, face, head, gestures; qualities of movement, like fast-slow, small-large, smooth-jerky, etc. In that study friends are asked to observe subjects and rate their behavior styles (along 78 items) and personality traits (24 items). Gallaher performed a factor analysis was performed on the two sets of items. Four dimensions were retained: ‘expressiveness’, ‘animation’, ‘coordination’ and ‘expansiveness’. ‘Expressiveness’ embodies factors such as quantity and variation of movements; ‘animation’ is associated with adjectives such as ‘lethargic-animated’ and can be related to tempo-velocity; ‘coordination’ is correlated with ‘jerkiness-smoothness’ and ‘awkwardness-gracefulness’; and ‘expansiveness’ is linked to the quantity of space. The person's behavior tendency was shown to be an innate individual characteristic that the author claimed to be a personality trait. In the second study she investigated the consistency of a person's behavior across time and situations. Results demonstrated this consistency: people that are quick when writing have a tendency to be quick while eating; if a person produces wide gestures then she also walks with large steps. Energy of movements is also an enduring characteristic, constant over time. This shows that human movements are consistent and irrespective of actions what they do.

2.3.2 Expressivity – Computer Scientists Views

2.3.2.1 Understanding Human Motion

Gestures can be acquired using electromagnetic sensors coupled with instrumented gloves to acquire gesture. In kinesiology the goal is to develop human body models and explain its mechanical functions and how one might increase its movement efficiency. A typical procedure involves obtaining 3-D joint data, performing kinematic analysis, and computing the corresponding forces and torques for a movement of interest (Calvert & Chapman, 1994). 3-D data is typically obtained in an intrusive manner, e.g., by placing markers on the human body.

Polana and Nelson (1994) referred to "getting your man without finding his body parts." Models for human action are then described in statistical terms derived from these low-level features or by simple heuristics. The approach without explicit shape models has been especially popular for applications of hand pose estimation in sign language recognition and gesture-based dialogue management.

The character of gestures during communication and the meaning it conveys are addressed by Zhao's (2001). He approached the gesture analysis on the basis of LMA and its effort and shape components. He used effort qualities and their combinations as a set of higher level features to be extracted for feature acquisition. For each effort dimension, dedicated neural network is constructed. Each network is trained, validated and tested over a number of motion samples. His research result says all the trained networks have a demonstrated accuracy of about 90 % in recognizing effort motion qualities for a group of people who deliberately made these expressions.

Camurri *et al.* studies (2003) illustrate analysis and classification of expressive gesture in human full-body movement and in particular in dance performances. They aim at (i) individuating motion cues involved in conveying the dancer's expressive intentions (ii) measuring and analyzing them in order to classify dance gestures in term of basic emotions, (iii) testing a collection of models and algorithms for analysis of such expressive content by comparing their performances with spectators' ratings of the same dance fragments. They used 4 layered approaches to model human movement and gesture, from low-level physical measures. Their analysis raised the hypothesis that the emotion categories can be differentiated by movement cues and that they are in line with the main predicted associations between emotion and movement cues.

For interpreting hand gestures, Quek (1995) used shape and motion features. According to Quek, during the hand motion individual fingers don’t contribute for hand gesture interpretation in gross motion. On the other hand, gestures in which fingers move with respect to each other will be performed with little hand motion.

2.3.2.2 Style and Content Separation by Low-Dimensional Mapping

To our knowledge, Tenanbaum and Freeman (2000) were the first who addressed style and content separation. They considered handwriting style, recognizing a familiar face or object seen under unfamiliar viewing conditions and familiar words spoken in unfamiliar accent. They model the mapping from style and content parameters to observations as a bilinear mapping. The approach is also related to the “learning-to-learn” (Thrun & Pratt, 1998) research program also known as task transfer or multitask learning. Defining “learning-to-learn” is that learning problems often come in clusters of related tasks, and thus learners may automatically acquire useful biases for a novel learning task by training on many related ones. Tenanbaum and Freeman work focuses on how learners can exploit the structure in families of related observations, bound together by their style and content interaction.

Tenanbaum and Freeman model the mapping from style and content parameters to observations as a bilinear mapping. They used two kinds of bilinear models in their research namely symmetric bilinear model and asymmetric bilinear model.

In the symmetric model, they represent style and content as vectors in observation space. In the observation space, weight function acts as a bilinear map from style and content to observation space. The symmetric model exactly reproduces the observation when the dimensionalities of style and content are equal. The model provides coarser but more compact representations as these dimensionalities are decreased.

Sometimes linear combinations of a few basis styles learned during training may not describe new styles. So it leads to go for asymmetrical bilinear model. The asymmetric model’s high-dimensional matrix representation of style may be too flexible in adapting to data in new styles and cannot support translation tasks.

Even though they define these bilinear models for the style and content separation in four different applications as we said earlier, our interest lies in handwriting style and content separation. They used different style fonts for the separation of style and content. They showed it is possible to learn the style of a font from observations and extrapolate that style to unseen letters, using a bilinear model.

In a nutshell, Tenanbaum and Freeman work show style and content (alphabets) can separate from hand written characters based on a mathematical model (bilinear model) and it needs to be further improved in terms of high dimensionality and a priori constraints in model parameters.

2.3.2.3 Stylistic Motion Synthesis from Human Motion

The problem of stylistic motion synthesis is addressed by Brand and Hertzmann (2000) through learning motion patterns from a highly varied set of motion capture sequences. Learning identifies common choreographic elements across sequences, the different styles in which each element is performed. The learned model can synthesize novel motion data in any interpolation or extrapolation of styles.

Style is varied in the mapping from qualitative states to quantitative observations. They use Hidden Markov Model (HMM), state space representation to make style distinction. State space representation has a Gaussian distribution over a small space of full body poses and motion which is added to the HMM a multidimensional style that can be used to vary Gaussian parameters resulting in stylistic HMM called "Style Machine".

They encoded style specific HMM by state means, square root covariance's and state dwell times. New styles are generated by interpolation and extrapolation within the space. The dimensionality of the space is reduced by Principal Component Analysis (PCA), treating each HMM as a single observation and the generic HMM as the origin. Typically, only a few stylistic degrees of freedom (DOFs) are needed to span the many variations in the training set and these become the dimensions of the style.

The algorithm automatically segments the data, identifies primitive motion cycles, learns transitions between primitives and identifies the stylist DOFs that make primitives look quite different in different motion capture sequences. Style machines are generative probabilistic models that can synthesize data in a broad variety of styles, interpolating and extrapolating stylistic variations learned from a training set which is shown in Figure 2-1. This work gives us an idea how stylistic motion can be synthesized in different varieties captured from real human motion.

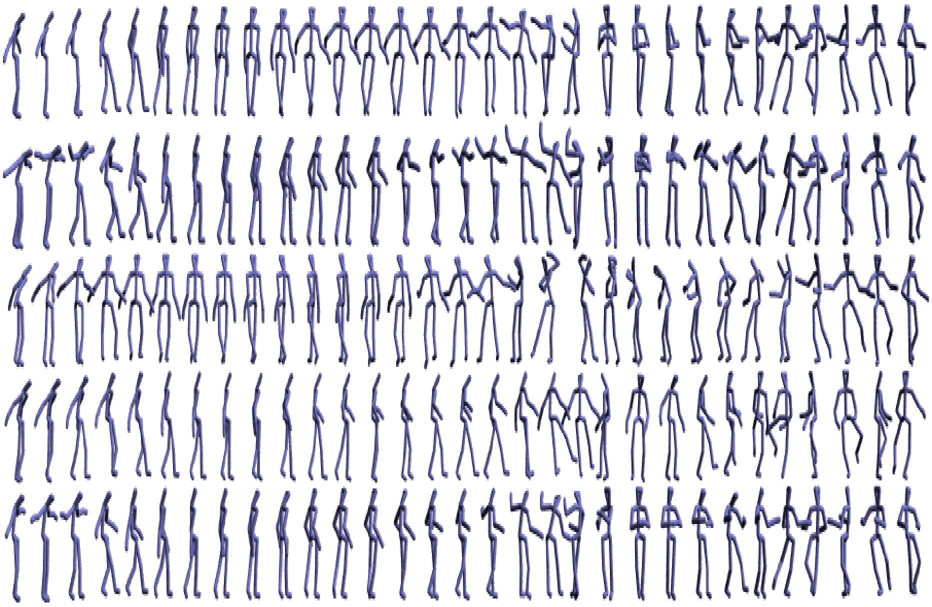


Figure 2-1 Different styles synthesized from the motion sequences. Five motion sequences synthesized from the same choreography, but in different styles(one per row). The actions, aligned vertically are tiptoeing, turning, kicking and spinning. The odd body geometry reflects marker placements in the training motion capture (Brand & Hertzmann, 2000).

2.3.2.4 Style and Content Separation using Dimensionality Reduction

Later, Elgammal and Lee (2004) proposed to separate style and content through non-linear dimensionality reduction. Elgammal and Lee described human motion such as gait, facial expression and gesturing, in nonlinear manifolds. They aim at separating style and content on manifolds representing dynamic objects from human silhouette through the walking cycle for their experiments. Given several sequences of walking silhouettes, with different people walking decomposing the intrinsic body configuration through the action (content) from the appearance of the person performing the action (style).

According to Elgammal and Lee, style is time-invariant personal feature and content is a representation of the intrinsic body configuration characterizes activity. They represent content in continuous domain and style is represented by discrete style classes. They adapted the Local Linear Embedding framework, the neighborhood of each point is determined by its nearest neighbors based on the distance in the input space and their objective is to find such weights that minimize global reconstruction error.

They used three dimensional embedding, since this is the least dimensional embedding that can discriminate the different body poses through the cycle. They used nonlinear dimensionality reduction approaches to capture the manifold geometry. After determining various people manifolds, they used LLE to embed different people manifolds. Given learned nonlinear mapping coefficients, the style parameters can be decomposed by fitting an asymmetric bilinear model to the coefficients.

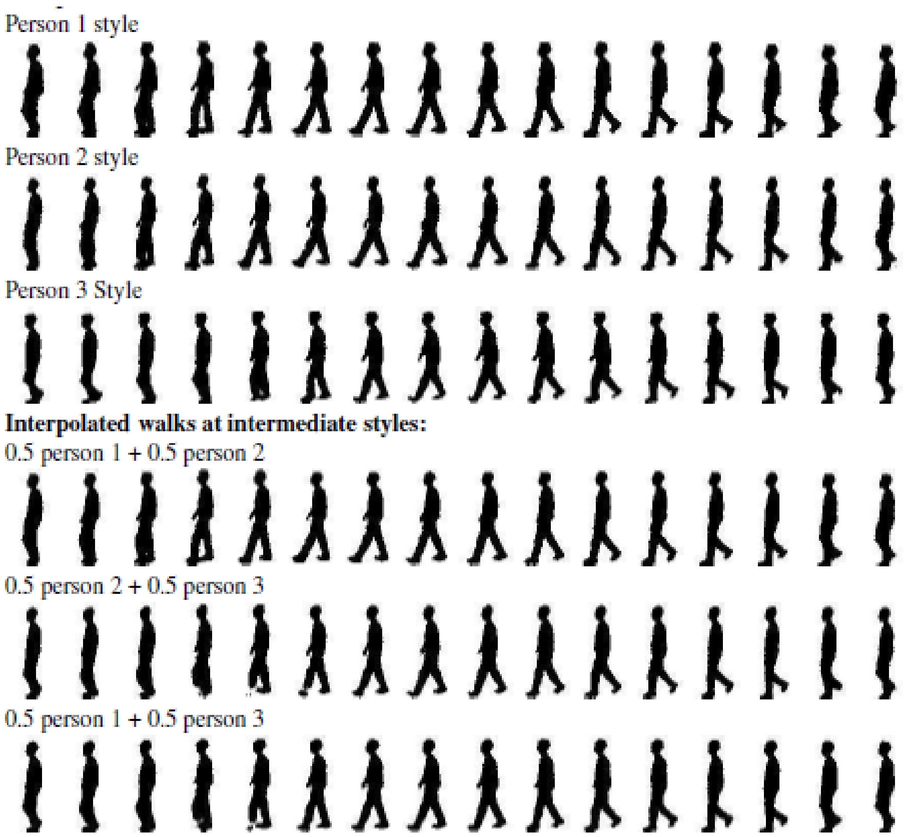

Person 1 style
Person 2 style
Person 3 Style
Interpolated walks at intermediate styles:
0.5 person 1 + 0.5 person 2
0.5 person 2 + 0.5 person 3
0.5 person 1 + 0.5 person 3

Figure 2-2 Interpolating styles and contents (Elgammal & Lee, 2004)

Elgammal and Lee learnt a decomposable generative model that explicitly decomposes the content as a function of time from the appearance (style) of the person performing the action as time invariant parameter. This framework decomposes style parameters in the space of nonlinear functions that maps between a learned unified nonlinear embedding of multiple content manifolds and the visual input space.

The lesson learnt from this work is, based on silhouette analysis we can separate style and content on manifolds representing dynamic objects. Elgammal and Lee create styles from different style (Figure 2-2.) This work addresses style and content separation for repetitive actions by combining a set of base styles like walking. It doesn’t address non repetitive gestures such as communicative gestures.

Urtasun *et al.* (2004) defined style based motion synthesis. They modeled the specific style of an individual whose motion had not yet been recorded through linear sums of principal components. They have presented a real time motion generation technique that allows them to generate the motion of a particular individual performing parameterized displacement activities. More specifically, they have investigated the case of walking, running and jumping. The first two are cyclical and parameterized by speed. The third one, a jump is noncyclical and parameterized in terms of its jump length. Given one single example, they modified the speed, length or body size while preserving the individual's specific style.

Instead of capturing full motion for each time for a person, in this work (Urtasun, Glardon, Boulic, Thalmann, & Fua, 2004) they extend the Principal Component Analysis (PCA) approach so that motion capture is drastically faster. In the whole classes of cyclic and non cyclic motions such as walking, running or jumping, it is enough to observe a person moving only once at a particular speed or jumping particular distance using either an optical motion captures system or a simple pair of synchronized video cameras.

They generate motion as a two step process. First, they project the new captured motion into the PCA space and measure its Mahalanobis distance to each recorded motion corresponding to the same speed or jump length. The generated motion is then taken to be a weighted average of motions at the target speed with the weights being inversely proportional to those distances. The principle for generating stylistic motion is identical. The new stylized motion is projected into the motion database. Weights are then computed based on the Mahalanobis distance and used to create the same style at a different speed. Figure 2-3 depicts a sneaking motion at 7 km/h generated by using a single example at 4.5 km/h and the standard walking database.

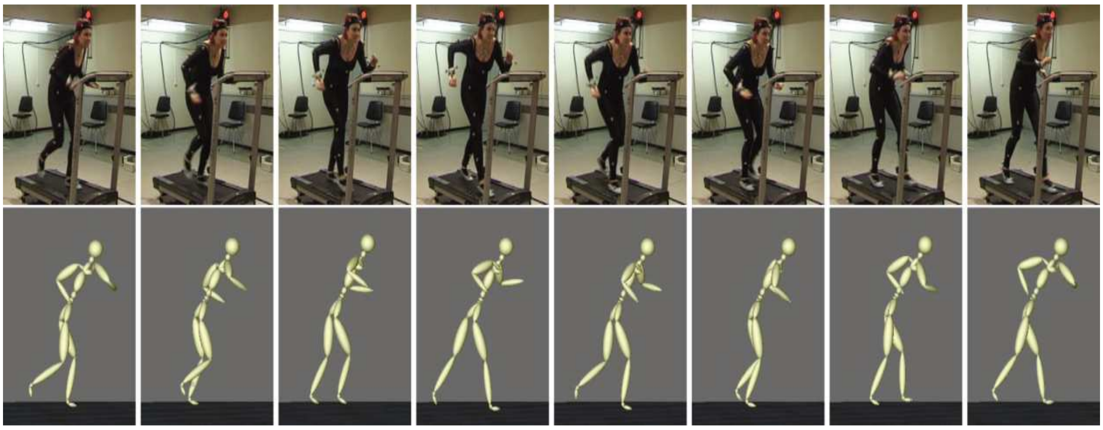


Figure 2-3 Generation of Stylized motion (Urtasun, Glardon, Boulic, Thalmann, & Fua, 2004)

In this method they have validated only three specific cyclic and non cyclic motions. Other kinds of motions are not experimented. This system uses motions that can be captured on a treadmill like straight line motion sequences other motions such as curvilinear motions are not considered since it is not captured through treadmill. The motions they considered are linear motions, for non linear motions PCA approach is not suitable.

Wang *et al.* (2007) used multi-factor Gaussian process models for style and content separation in human motion. They developed multi linear models using non linear basis functions for human locomotion data in which each pose is generated by factors representing the person's identity, gait and the current state of the motion. They applied the multifactor model to human motion data consisting of sequences of poses. A single pose is represented as a feature of 89 dimensions, including 43 angular DOFs their velocities, and the global translational velocity. For any particular motion sequence, it is assumed that style stays constant over time. They approached this problem in two ways namely Gaussian Process dynamics and Circle Dynamics Model (CDM).

They do not constrain corresponding poses in different sequences to share the same content because prior knowledge of the exact correspondences is not assumed. It is desirable, to restrict the content of different styles to lie on the same trajectory, especially for motion synthesis. Wang *et al.* (2007) demonstrates the ability of the model to generate smooth transitions from walking to running and from running to striding. The transitions are generated by linearly interpolating the gait vector with respect to the changing state vector. The subject vector is fixed to a particular person.

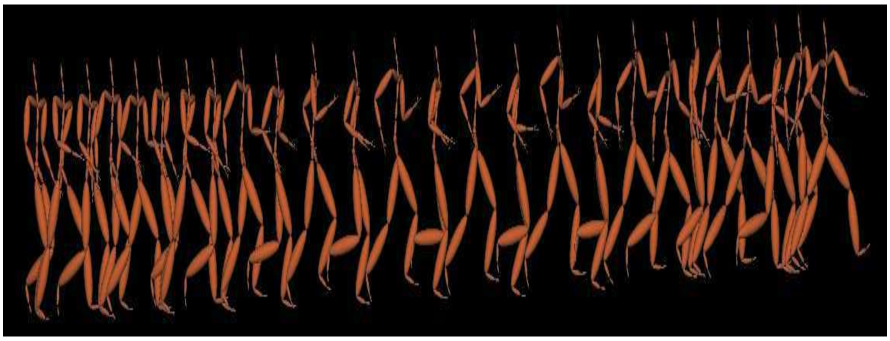


Figure 2-4 Transitions between different motions are achieved by linear interpolation (Wang, Fleet, & Hertzmann, 2007)

Later Wang *et al.* (2008) analyzed the human motion through non linear analysis. They used Gaussian Process Dynamic Models (GPDM) for non linear analysis. They learn models of human pose and motion from 50 dimensional motion capture data using GPDM. GPDM represents 50 dimensional motion capture data in low dimensional latent space with associated dynamics, as well as map from the latent space to observation space. This results in a non parametric model for dynamical systems that account for uncertainty of the model. Despite the use of small data sets, the GPDM learns an effective representation of the non linear dynamics in the latent space.

2.3.2.5 Separating Style based on Time Invariant Model

Hsu *et al*. (2005) proposed to transform the style of human motion while preserving its original content, by aligning motions and linear time invariant model (LTI) to represent stylistic differences. For two motions in different styles typically contain different poses, Hsu *et al* propose iterative motion warping, which automatically computes dense correspondences between stylistically different motions. Motion warping demonstrates that many variations in motion can be modeled by smooth spatial distortions of individual degrees of freedom.

This method uses a LTI model to approximate the relationship between the input and output styles. Once the parameters are estimated from the training data, the model translates new motion with simple linear transformations. The output retains the content and the detail of the original input, but differs in its style of execution.

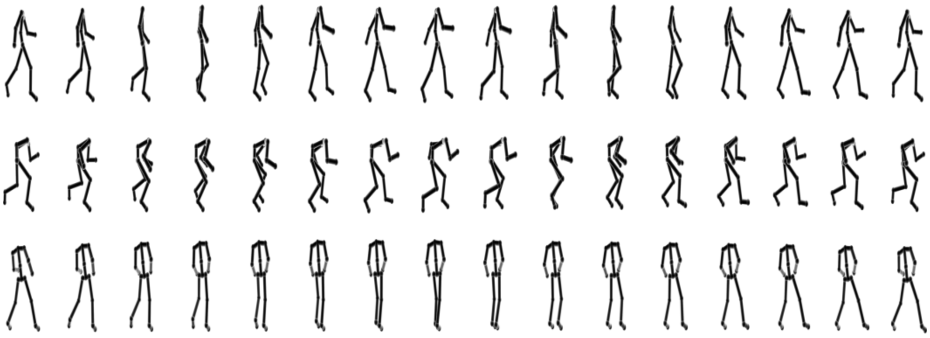


Figure 2-5 Style translation system transforms a normal walk (TOP) into a sneaky crouch (MIDDLE) and a sideways shuffle (BOTTOM) (Hsu, Pulli, & Popović, 2005)

They conducted some experiments using this algorithm. The data set contained various stylized locomotion and each style is performed in slow, medium and fast speeds. The results of this algorithm, iterative motion warping gave consistently better results which are shown in Figure 2-5. The coordinates of the motions were then interpolated and extrapolated to allow for continuous control of style and speed parameters.

Human identification based on gait analysis has been widely studied (Lee & Grimson, 2002) (Chellappa, Roy-Chowdhury, & Kale, 2007) (Ekinci, 2006) Gait, which can be regarded as a walking style, does convey information on the identity of a person.

Mancini *et al.* (2007) map acoustic cues and emotion to expressivity parameters they use to control the expressive virtual head of the Greta Embodied Conversational Agent (ECA). Pelachaud and Poggi (2002) aim at combining the Greta facial expressions in a complex and subtle way, just like humans do, by assessing and managing the multimodal communicative behavior of a person when different communicative functions have to be displayed at the same time. Pollick (2003) points out limits of dissecting movement features and ascribing discrete values to movement frequency and speed: the degree and manner in which this style is dependent on spatial and temporal encoding is not trivial and varies between different movements.

2.3.2.6 On-line Motion Style Transfer

Wu *et al*. (2006) propose a fast and convenient algorithm for human-motion style editing. They define the style of a motion as statistic properties of mean and standard covariance of joint quaternion in 4D unit sphere space. Their algorithm can transfer the style of a motion to another by transferring these properties.

The goal of their approach is to transfer the style of the reference motion M_R to the source motion M_S, producing a target motion $M_{T.}$ The target motion can preserve the details of the source motion and can inherit the style of the reference motion. Their approach is shown in Figure 2-6.

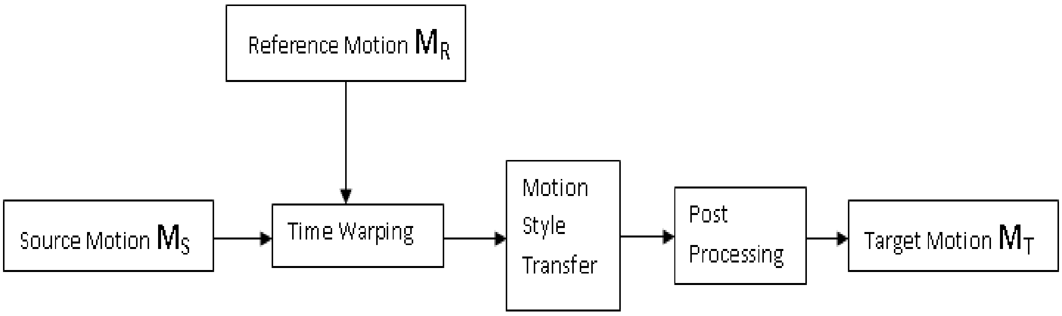


Figure 2-6 Wu *et al*. (2006) Approach

M_S and M_R are input directly into the time warping module to setup frame correspondence. Then, the time-warped motions are fed into the statistic style transfer module for style transferring. In the post-processing stage, they apply a reverse time warping to obtain target motion M_T. The obtained target motion will have the style of the reference motion and preserve the details of the source motion.

They define the style based on Reighar *et al.* (2001) work. Reighar *et al.* define the style of an image as the mean and standard covariance of color components in a linearized space, and has successfully transferred the style of an image to the other by transferring the mean and standard variance of color components in this space. Using this definition, Wu et al. transfer the motion style by modifying the mean and variance of the source motion M_S according to that of the reference motion M_R. They apply this algorithm in quaternion domain and obtained smooth results. Their results are shown in Figure 2-7.

From this work, we see that style is defined based on image processing approach and it is applied to the full body motion. Also this approach cannot be applied to transfer style between figures that do not have identical structures. The style is defined through low level parameters joint angles.

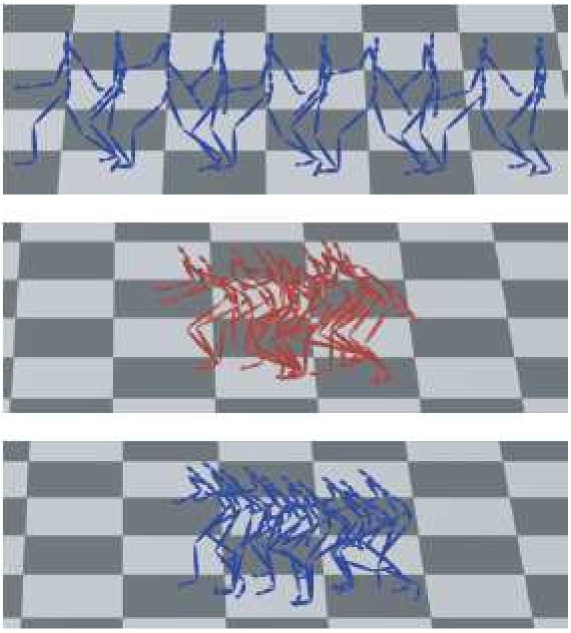


Figure 2-7 Transfer the "stealthy" style of a walking motion (middle) to another walking motion (top) produces a new "stealthy walking" motion (bottom).

2.3.2.7 Expressive Animation

We start our discussion with how complex human motion (e.g., dancing) is synthesized and then proceed on with available methods to animate the human motions in virtual environment. In the end we will finish up with how these animation methods bring style and emotion in the virtual character.

There are many works aiming at generating modified animations based on motion capture data. Li *et al.* (2002) propose a technique called motion texture for synthesizing complex human figure motion that is statistically similar to the original motion capture data. They describe motion texture as a set of motion textons[2] and their distribution which characterize the stochastic and dynamic nature of the captured motion. They define motion sequence into segments, such that each segment can be represented by one of the textons. Multiple textons could be represented by the same texton. A motion texton is represented by a linear dynamic system (LDS) that captures the dynamics shared by all instances of this texton in the motion sequence. Once the motion texture is learnt, it can be used for synthesizing novel motion sequences. The synthesized motion is statistically similar to, yet visually different from, the motion captured data. This model enables users to synthesize and edit the motion at both the texton level and the distribution level. Synthesized motion is shown in Figure 2-8.

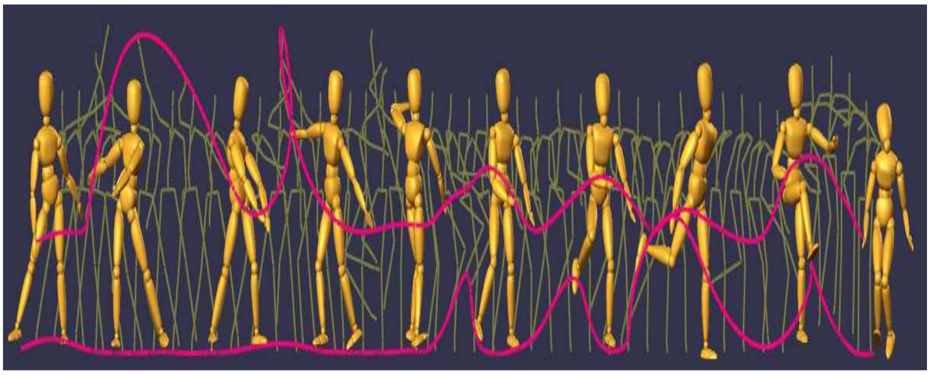


Figure 2-8 This 320-frame sequence of dance motion is choreographed from (1) the starting frame, (2) the ending frame and (3) the learnt motion texture from motion captured dance data. Four motion textons are generated from the motion texture and then used to synthesize all the frames in this sequence. A number of key frames are also shown in the figure to demonstrate that the synthesized motion is natural, smooth and realistic (Two red lines indicate the trajectories of the right hand and right foot) (Li, Wang, & Shum, 2002).

Kopp *et al* (2003) explained a virtual agent can imitate and recognize natural gestures performed by a human using marker motion capture. The imitations have two phases. In the first phase, the agent extracts and reproduces the essential form features of the stroke which is the most important gesture phase. The second phase is the meaning based imitation level that extracts the semantic content of gestures to re express them with different movements.

Arikan and Forsyth (2002) developed a method for automatic motion generation at interactive rates. They establish high level constraints and a random search

[2] Textons refer to fundamental micro structure composed of local image features , are “the putative units of pre-attentive human texture perception” (Julesz, 1981).

algorithm is used to find the right pieces of motion data to fill in between. Close to their work, the concept of a motion graph (Kovar, Gleicher, & Pighin, 2002) is defined to enable one to control a character's locomotion. The motion graph contains the original motion and automatic generated translations and allows the user to have high level control on movements of the characters.

Lee *et al.* (2002) developed a new technique to control a character in real time using possible interfaces. They aim at obtaining a rich set of avatar behaviors is to collect an extended, unlabeled sequence of motion data appropriate to the application. So they developed a method for such a motion database and it's preprocessed for flexibility in behavior and efficient search and exploited for real-time avatar control. Three interfacing techniques are used to control avatar motion using the data structure videlicet user select set of available choices, user sketches a path through an environment or acts out a desired motion in front of a video camera. All three interfaces are tested and shown in Figure 2-9.

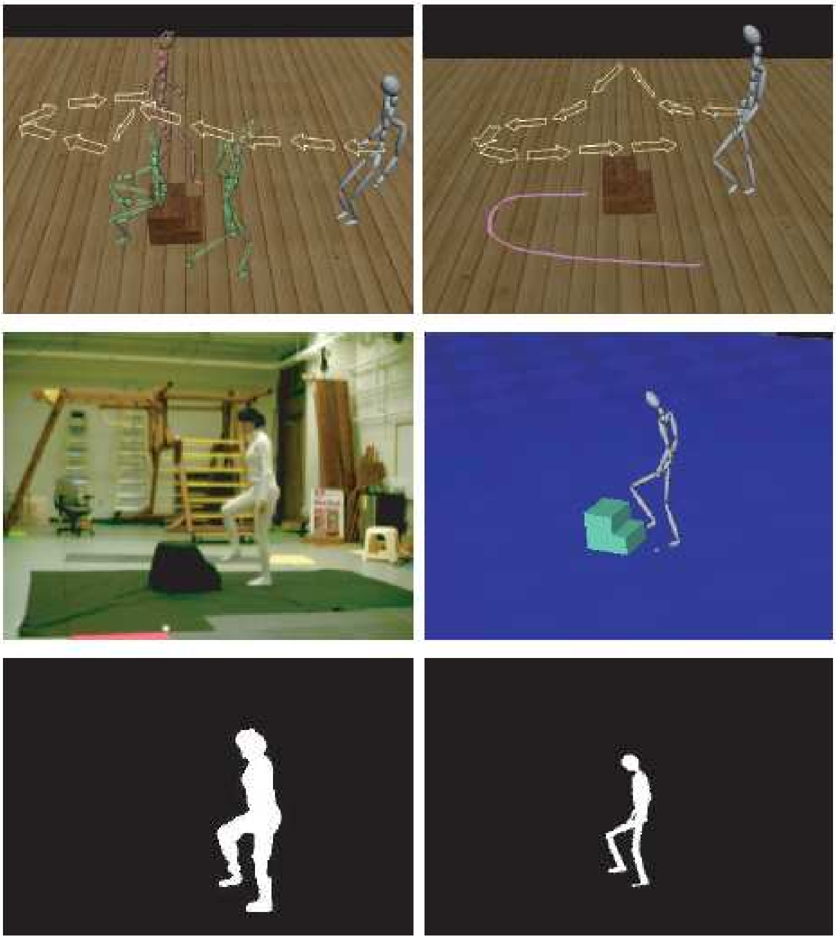


Figure 2-9(Top Left) Choice based interface. (Top Right) Sketch based Interface. (Middle Left and Bottom Left) The user performing a motion in front of a video camera and her silhouette extracted from the video. (Middle Right and Bottom Right) the avatar being controlled through the vision based interface and the rendered silhouette that matches the user's interface (Lee, Chai, Reitsma, Hodgins, & Pollard, 2002)

This method gives us an introduction, how the real human motion can be brought in the virtual environment as personalized avatars. In this work, they showed how to handle the motion capture data in terms of human behavior. This work will be the starting point in deriving identity and mood of a real human in virtual environment.

From the results, sketch based interface is easy to use and responds instantly to the user input by finding a matching action and updating the display while the path is being drawn. Sometimes, the sketch-based interface was not able to distinguish different styles. In the case of vision based interface, the silhouette comparison usually can discriminate different styles. If the user acts out a motion that is not available in the database, the system selects motion that looks similar.

Any of the above techniques would be appropriate when the user has a large database of movements.

The emotional state of a person can be recognized from facial expressions. It is a combination of sadness, anger, worrying, uncertainty, happiness and surprise (Poggi & Pelachaud, 2000). Pelachaud (2009) developed a model that is based on perceptual studies and encompasses several parameters that modulate multimodal behaviors. This emotional state of the person is also carried out based on bodily motion (Niewiadomski, Hyniewska, & Pelachaud, 2009).

In the process of animating quality human movements, Chi *et al.* (Chi, Costa, Zhao, & Badler, 2000) proposes a method that allows animators to enhance the style of pre existing motions in a natural way. They use the principles of Laban Movement Analysis to create a new interface for the application of the particular qualities movement to movement.

Ball and Breese (2000) work gives correlation between temporal and spatial tendencies in gesture/posture and personality/emotion – movement frequency and speed were related to emotional arousal, as was the size of overall body outline.

We started to discuss how human motion is brought in the virtual environment and within the next section we address how the identity and mood of a real human can be brought in a virtual agent.

2.4 Expressivity Parameters

Computer scientists (Hartmann, Mancini, & Pelachaud, 2005) (Poggi, 2001) designed bodily expressive virtual agent based on the psychologists study. This helps us to design the virtual agent with high level parameters. They present a computational model of gesture quality. Once a certain gesture has been chosen for execution they characterize bodily expressivity with a small set of dimensions derived from a review of psychology literature (Wallbott, 1998).

Based on the introspection and observational studies Poggi (2001) proposed to analyze the gesture movement based on following parameters.

a) Part of the hand or arm involved – different parts of the limbs may be involved in the gesture movement. For example, saying “no” by shaking only the index finger with still fist is less intense than shaking it along with the whole fist, or even with the entire forearm.
b) Direction – the point in the space toward which the gesture is directed. Forward, backward, outward, inward, upward, downward and their combinations.
c) Path – the route of a gesture outlines in space (straight, oblique, circular, half-circular, thrumming, oscillation, etc.,)
d) Size – how large is the movement in width (long, short, narrow)
e) Pressure – the strength of the movement, which includes three sub parameters:
    1. Tension – the muscular tension of the hand or arm in performing the movement (tense, normal relaxed)
    2. Impact – the way in which the gesture stops at the end of the movement (block, normal, skim)
f) Tempo – the set of temporal features of the movement, that can be distinguished in to three categories namely
    1. Duration - how long the movement lasts (long, normal, short)
    2. Speed – how speedy the gesture moves (fast, slow, normal)
    3. Rhythm – if and how the gesture is repeated and in which rhythmical structure.

Hartmann *et al.,* (2005) characterize expressivity for human bodily movement based on perceptual studies conducted by Wallbott (1998) (1985), Wallbott and Scherer (1986) and by Gallaher (1992). Their approach is driven by a perceptual stand point – how expressivity is perceived by others. So their work focuses on surface realizations of movement and do not attempt to model underlying muscle activation patterns. They define an intermediate level of behavior parameter as a useful enabling tool to facilitate the mapping of holistic, qualitative communicative functions such as mood, personality and emotion to low level animation parameters like joint angles. On the basis of these literatures Hartmann *et al.,* (2005) proposed six dimensions that characterize expressivity in qualitative terms. They are:

a) Spatial Extent – amplitude of movements. That is, quantity of space taken by a body part (how extended are the arms). This dimension is related to the dimension ‘expansiveness/spatial extension’ defined by Wallbott and to the ‘dimension expansiveness’ by Gallaher.
b) Temporal Extent – Velocity of execution of a movement t (how fast or how slow an arm moves). It is related to the ‘animation’ factor of Gallaher.
c) Power – Dynamic properties of the movement m (weak vs. strong). It is related to the degree of acceleration of body parts. It corresponds to the dimension ‘movement dynamics/energy/power’ defined by Wallbott.
d) Fluidity – Level of continuity of successive movements (jerky *vs.* smooth movements). It is similar to the ‘coordination’ dimension defined by Gallaher.
e) Overall Activation – Overall quantity of movement on a given channel for the whole animation (many hand gestures *vs.* none, passive / static or animated / engaged). This dimension embodies similar information as the ‘expressiveness’ dimension defined by Gallaher.
f) Repetition – Tendency to rhythmic repeats of specific movements. This dimension is newly added.

Each of the parameters except repetition is float-valued and defined over the interval
[-1, 1], where zero point corresponds to the action without expressivity control. Repetition parameter has the values of 0, 1 and 2.

2.4.1 Related works on Expressivity Parameters

Castellano and Mancini (2009) analyzed the emotional gestures through expressivity parameters and animated the emotional gestures in an embodied agent. Their system allows for the real-time analysis of human movement and gesture expressivity and the generation of expressive copying behavior in an agent which is shown in Figure 2-10. Their system two different platforms: EyesWeb (2004) and Greta.

EyesWeb is used track the human motion and EyesWeb Expressive Gesture Processing Library is used to extract the expressive motion cues. Similar to the expressivity parameters (Hartmann, Mancini, & Pelachaud, 2005) they used contraction index (CI), velocity, acceleration and Fluidity based on Wallbott’s work (1998).

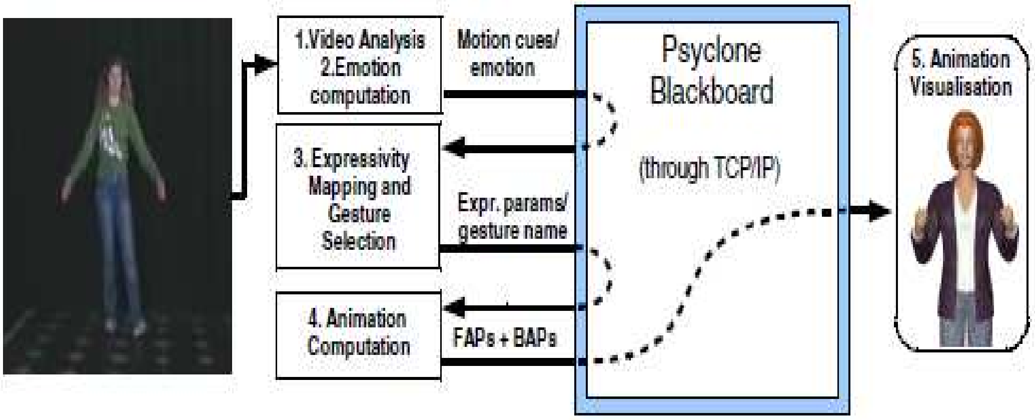


Figure 2-10 Architecture of Castellano and Mancini's (Castellano & Mancini, 2009) system.

Human motion is tracked continuously through EyesWeb. They do not segment the human movements but they segment the gestures based on time window. They define a time window at the end of which they send the computed motion cues from EyesWeb to the rest of the system. The duration of the time window corresponds to the duration of a gesture performed by the agent.

After extracting expressive motion cues in real time, their system associates them with emotions joy, anger or sadness. Based on Wallbott's study (1998) they defined correspondences between expressive motion cues and emotions. They do linear mapping between contraction index, velocity, acceleration, fluidity and spatial extent, temporal extent, power, fluidity of Hartmann's *et al.* (2005) expressivity parameters. With the mapped values as input to the Greta, the virtual agent is animated.

They conducted perceptual study to evaluate their estimation. In their test, they evaluated the three different emotions, joy, anger and sadness. They selected videos from GEMEP corpus (Bänziger, Pirker, & Scherer, 2006), in which two persons showing these emotions separately, totally 6 videos are selected for perceptual study. In their study, twelve users are asked to observe the animated movements and asked to choose an emotional label (joy, anger or sadness). User's emotion recognition rated 83.3 % for joy and anger and 70.8 % for sadness for the Greta animated virtual agent.

In their work, the EyesWeb parameters (CI, velocity, acceleration and fluidity) are evaluated for the discrete motions. Discrete motions are made based on time window, not based on gesture segmentation. They aim at replicating human emotions such as joy, anger and sadness in the virtual agent. This work doesn't concern about the style of the person.

Caridakis *et al.* (2008) work is about multimodal and expressive synthesis on virtual agents, based on the analysis of actions performed by human users. They consider image sequence of recorded human behavior as input. They analyze both facial and hand gestural aspects of the user's behavior for the multimodality approach. We are interested in their hand gestures mimicking.

Hand motion tracking is starts with identifying skin color from the input image sequence. They create moving skin masks and skin color areas are tracked between subsequent frames. By tracking centroid of those skin masks, they produce estimation for user movements. The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions.

They use Hartmann *et al.* (2005) expressivity parameters because it tackles all the parameters of expression of emotion. They implemented five of Hartmann *et al.* (2005) parameters such as overall activation, spatial extent, temporal extent, fluidity and power.

They estimate overall activation from actor (real human who performs gestures) as sum of motion vector's norm. Spatial extent is calculated as the maximum Euclidean distance of the position of the hands of the actor. To extract fluidity from input image sequence, they calculate the sum of the variance of the norms of the motion vectors. Power is calculated from the derivative of the motion vectors.

They considered twelve actors gestures as input image sequence for their experiments. These actors performs single emotional gesture namely bored, wave, explain, clap, raise hand, and 'leave me alone', 'oh my god'.

Expressivity parameters are estimated from these emotions and fed as input to the Greta gesture engine to replicate the actor's gesture. Their architecture is shown in Figure 2-11.This work aim at replicating human actions with expressivity.

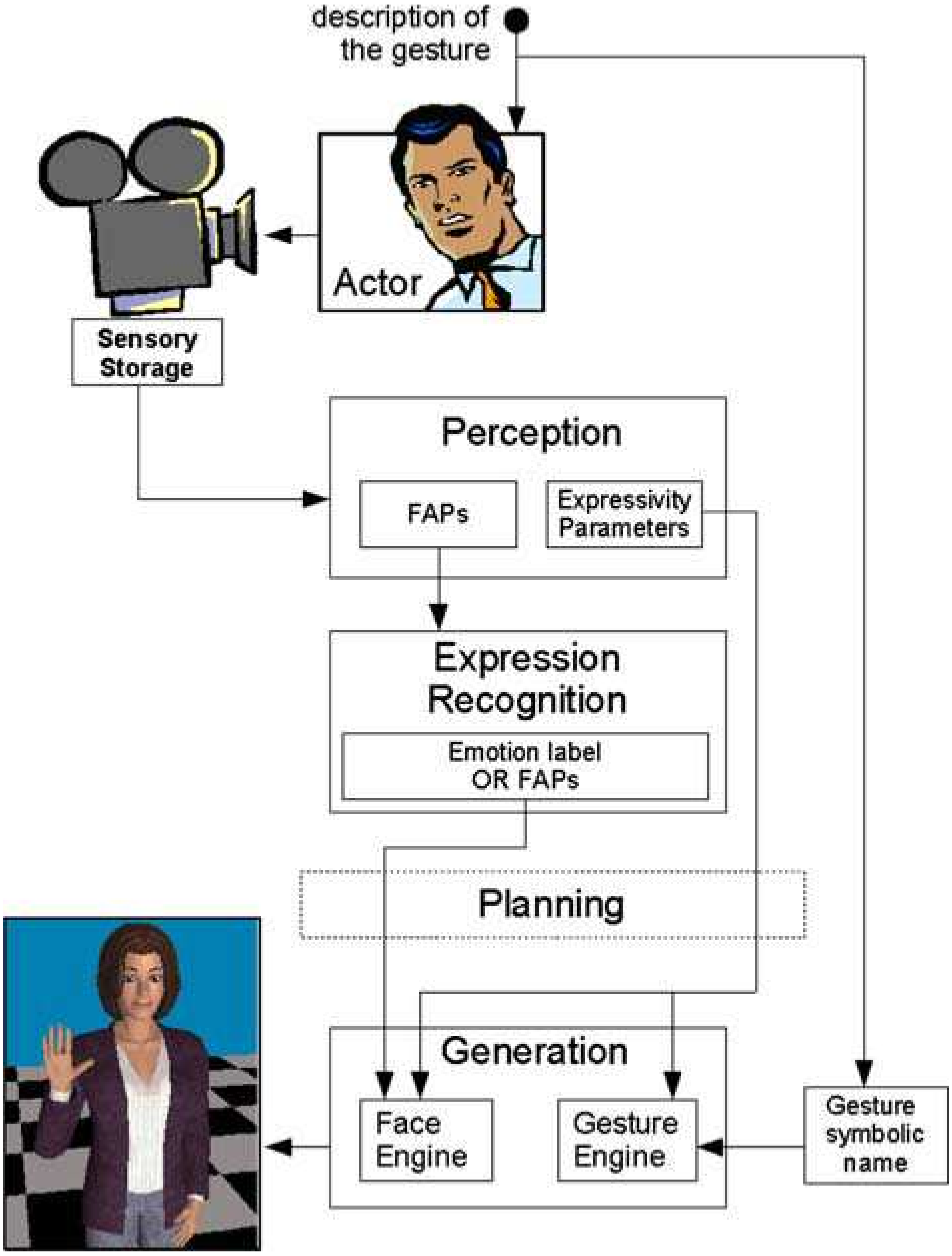


Figure 2-11 Architecture of Caridakis e*t al* (2008) system.

2.5 Customized Virtual Agents

A number of platforms (IMVU, Inc.) (Linden Research, Inc) allows us to animate an avatar, but only few virtual agents (Noot & Ruttkay, 2005) (Hartmann, Mancini, & Pelachaud, 2002) have high level interface for expressivity control. We will elaborate the architecture and its sample images in this section.

2.5.1 Virtual Agent Based on GESTYLE Language

Noot and Ruttkay (2005) proposed Gesturing Style (GESTYLE) language that aim at non verbal expression. It is a markup language to annotate text to be spoken by an ECA, to prescribe the usage of hand, head and facial gestures accompanying the speech in order to augment the communication.

GESTYLE language is written based on XML defines what gestures an ECA “knows”, and what are the habits of using these gestures, concerning intended meaning, modalities and subtle characteristics (like ethnicity and personality of the ECA) of the gestures. In their language high level tags are used to define style for ECA and the appropriate gestures to be performed. High level tags are profession (carpenter, surgeon, etc.) and culture (American culture, Japanese culture, etc.). GESTYLE is hierarchically organized as basic gestures and composite gestures. Composite gestures are formed by combing two or more basic gestures through gesture expressions. In the next level, the meanings denote the communicative acts which can be expressed by some gestures. A meaning is mapped to one or more gesture expressions, each specifying an alternative way to convey the same meaning. The mappings of meanings to alternatives of gestures are given as entries of style dictionaries. A style dictionary contains a collection of meanings pertinent to a certain style.

In GESTYLE language the style of the ECA is defined based on gesture expressions which are established on style dictionaries. The style dictionaries are at the core of GESTYLE: they are crucial in the specification of different styles. In a style dictionary, the characteristics are given for an individual: professional or cultural group, or people of certain age, sex or personality. The expressive gestures are predefined in GESTYLE based on style dictionaries. And this style dictionaries is not universal, which is to be modified based on our needs. Virtual agent generated through GESTYLE language is shown in Figure 2-12.

More importantly GESTYLE, XML based markup language and it is not implemented in real time. Even though its architecture is completely defined, the real time implementation and testing with real human is not done.

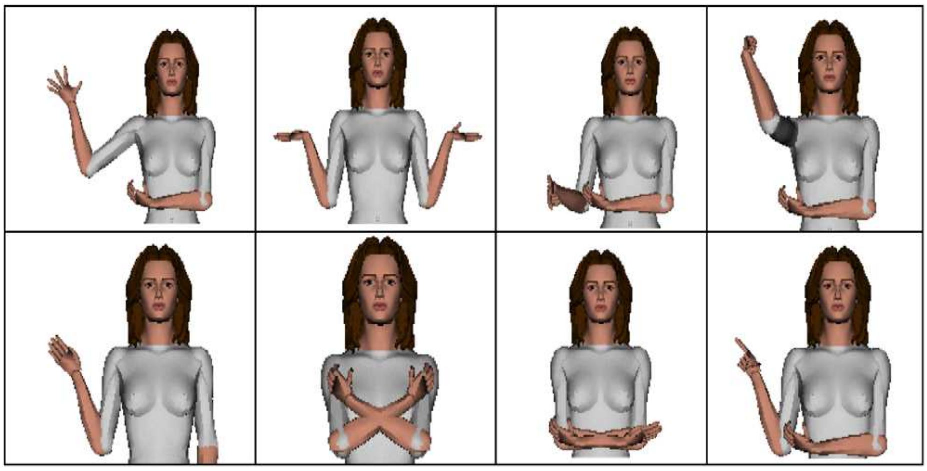


Figure 2-12 Sample stills of virtual agent which is generated using GESTYLE which are taken from a longer demonstrator for GESTYLE which renders in VRML and in the STEP system Eliens et al (2002)

2.5.2 Greta

Greta (Hartmann, Mancini, & Pelachaud, 2002) is an ECA which is able to communicate using a rich palette of verbal and nonverbal behaviors. Greta is real-time three dimensional female agents, compliant with MPEG-4 animation standard. She can talk and simultaneously show facial expressions, gestures, gaze, and head movements. Two standard XML languages Function Markup Language (FML) and Behavior Markup Language (BML) allow the user to define her communicative intentions and behaviors.

Greta's architecture is shown in Figure 2-13. The engine produce animation data in MPEG-4 compliant Face Animation Parameter (FAP)/ Body Animation Parameter (BAP) format which in turn drive a facial and skeletal body model in OpenGL. Gesture engine first performs text to speech conversion through 'Festival' (Black, Taylor, Caley, & Clark) which provides necessary phenomenon timing for synchronizing gesture to speech.

Communicative function tags which are candidates for gesture matching are extracted in 'Turn Planner'. The 'Gesture Planner' matches communicative function tags to a library of known prototype gestures and also schedules rest phases when arms are retracted to the body. The 'Motor Planner' then concretizes abstract gestures by calculating the key frame joint angles and timing, Finally, a bank of different 'Interpolators' generate in between frames to complete the animation.

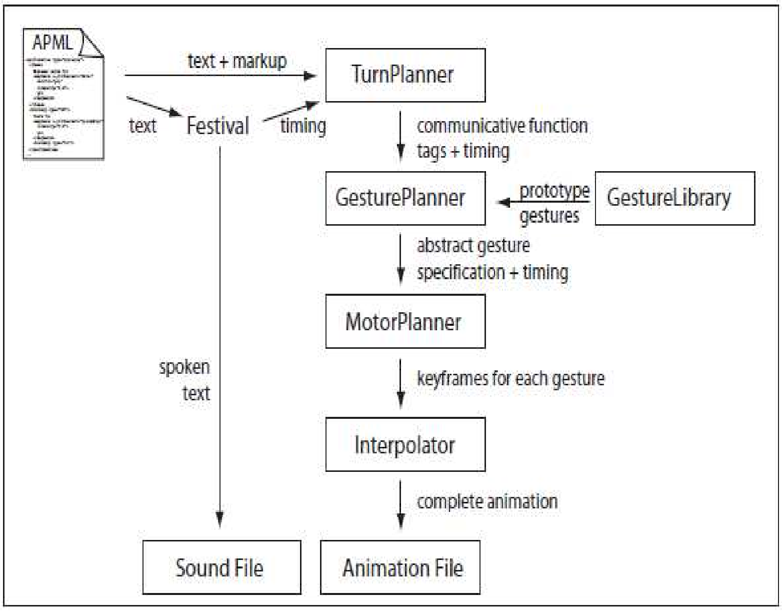


Figure 2-13 Greta Architecture outline (Hartmann, Mancini, & Pelachaud, 2005)

Using the Hartmann's *et al.* (2005) expressivity parameters, the intermediate poses are calculated from the key frames. Greta gestures are customized by the input expressivity parameters. Some examples of Greta face and body pose is shown in Figure 2-14. Even though there are six expressivity parameters we explain two basic Hartmann's *et al.* expressivity (2005) parameters namely spatial extent and temporal extent more in detail.

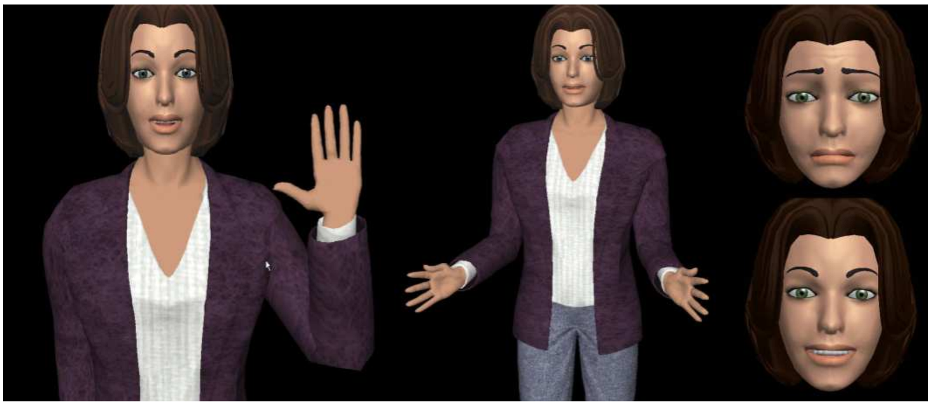


Figure 2-14 Examples of Greta gestures and facial expressions (Pelachaud, 2005)

2.5.2.1 Spatial Extent

Spatial extent describes the space used by a person for making gestures. Hartmann *et al.* define spatial extent as a parameter controlling the centers of the McNeill's sectors (1992) where hand movement occurs. Spatial extent value +1.0 (resp. −1.0) means the wrist moves further (resp. closer) to the coordinate origin, which is set at the sacroiliac vertebra. Minimum and maximum spatial extent is pictorially shown in Figure 2-15.

They replace wrist positions $p = (p_x, p_y, p_z)^T$ in the neutral spatial extent trajectory with positions $p' = (p'_x, p'_y, p'_z)^T$:

$$\begin{bmatrix} p'_x = (1 + SPC \cdot Agent_x)p_x \\ p'_y = (1 + SPC \cdot Agent_y)p_y \\ p'_z = (1 + SPC \cdot Agent_z)p_z \end{bmatrix} \tag{2.1}$$

where $\bullet_x$, $\bullet_y$ and $\bullet_z$ refer to the lateral, vertical and frontal directions. *SPC* represents spatial extent. *Agent*$\bullet$ are scaling factors in the respective directions used in the Greta animation engine (resp. 1.3, 0.6 and 0.25 resp. in the *x*, *y* and *z* directions and positive *SPC*, and resp. 0.7, 0.25 and 0.25 resp. in the *x*, *y* and *z* directions and negative *SPC*), and *SPC* is the spatial extent in the range [−1.0, +1.0].

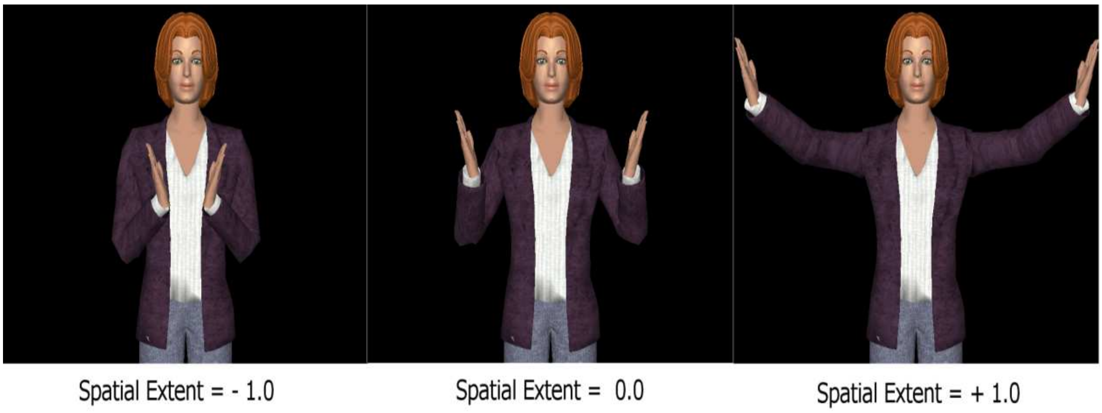


Figure 2-15 Spatial Extent hand variation , (left) minimum spatial extent gesture, (middle) neutral spatial extent gesture, (right) maximum spatial extent gesture.

2.5.2.2 Temporal Extent

Temporal extent (*TMP*) describes how fast a gesture is performed. According to Hartmann *et al.*, (2005) time for performing a gesture is segmented into three intervals, namely preparation, stroke and retraction based on Efran's (1941) theory. A stroke is that part of an expressive gesture that conveys meaning. It is preceded by a stroke preparation period when the wrist moves from the initial position, and followed by retraction where the wrist returns to the rest position after completing the stroke.

Temporal extent is related to the stroke time interval. Hartmann *et al.* derive the time taken for each segment from a simplification of Fitt's law (1954) :

$$T = a + b\log_2(\|x_n - x_{n+1}\| + 1) \tag{2.2}$$

where T is the duration of the stroke, a is a time offset, b is a velocity coefficient, x_n is the wrist positions at stroke start, x_{n+1} is the wrist positions at the stroke end. The velocity coefficient is defined by Hartmann *et al.* as:

$$b = \frac{(1 + 0.2\ \cdot TMP)}{10} \tag{2.3}$$

where *TMP* is the temporal extent where –1.0 TMP corresponds to lower speed and +1.0 TMP to higher speeds. The demonstration of minimum, neutral and maximum temporal extent values are shown in Figure 2-16. The variation of wrist in vertical direction with respect to shoulder is shown.

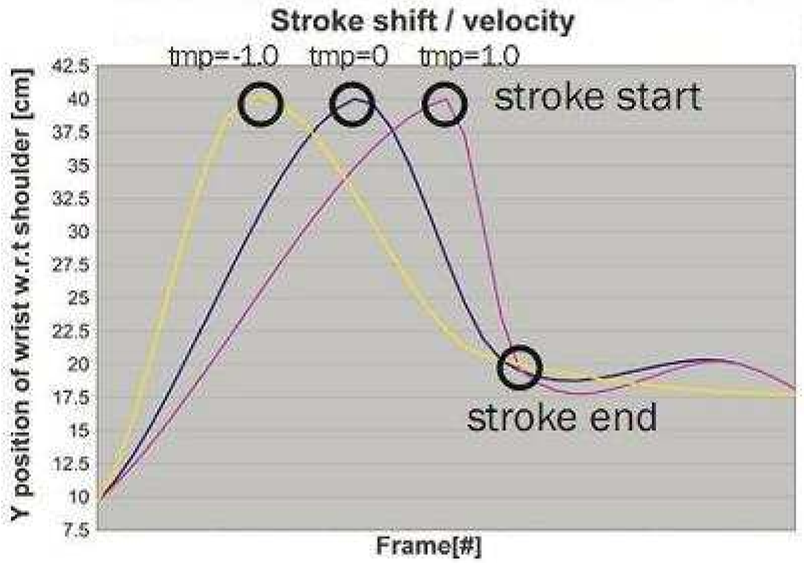


Figure 2-16 Temporal Extent - plot of wrist position over time in one dimension. (Hartmann, Mancini, & Pelachaud, 2005)

2.5.2.3 Repetition

To further increase the realism of gestural performance Hartmann *et al* (2002)introduced the concept stroke expansion. Observation of some human gesturing expresses additional rhythmical emphasis in a sentence. In these cases, the first execution of the stroke of a given gesture carries its usual semantic function; afterwards, however, the hand remains in its assumed shape and the arm partially repeats the gesture's stroke movement to further accentuate the rhythm of the associated speech. To measure the rhythmic a parameter is introduced as repetition, which measures the number of time stroke is repeated.

2.6 Conclusion

We aim at rendering virtually the gesturing style and mood of a real human. Psychologists state that gesture expressivity depends on emotions (Wallbott & Scherer, 1986). It can be described with the quantity of movement of hand and head (Wallbott, 1985). Focusing on hand communicative gestures, relevant motion features are hand motion having quantity and quality of movements that defines mood of a person. Mood is related to emotions, such as joy, sadness, fear, anger etc. that are common to all humans which is not unique to a human being. But the quality and quantity of hand movements during non verbal communication is consistent irrespective of actions and emotions for a person (Gallaher, 1992).

From our literature review about human's arm motion, shape and motion features contribute to gesture expressivity (section 2.3). Also hand configuration doesn't contribute to the expressivity of a gesture. Another way of analyzing arm motion is through gesturing style, defined mostly from body gestures but that also encompasses face expression.. Gesturing style is a feature related to identity of a person across activities (Gallaher, 1992). The mood of a person is a prolonged emotional state caused by the cumulative effect of momentary emotions and is a dynamic property that changes with time (Thalmann, Jain, & Ichalkaranje, New Advances in Virtual Humans: Artificial Intelligence Environment, 2008).We shouldn't confuse style, emotion and expressivity of a person. Expressivity encompasses both style and mood of a person. The information conveyed through gestures in a given situation may not be unique for human being, but the way it is being communicated is consistent, which we call it as style. Also information conveyed through gestures comprises mood of the person. We understand from our earlier literature survey that every human gestures with his town style and his / her mood.

We need to identify and define the style and mood of a person. We have seen that many approaches have been proposed to describe and identify the style and mood of a person. Approaches which are trying to capture the style analyze walking, running or jogging of a person. We are not aware of previous attempt to retrieve style as an identity feature of a person from communication gestures upper body or hand motions. In order to determine expressivity of a person in our work we use the expressivity parameters (Hartmann, Mancini, & Pelachaud, 2005) based on emotions of a person along with his natural consistent behavior. That is, the expressivity parameters are built on Gallaher (1992), Wallbott and Scherer (1986) and Wallbott's (1998) work.

Previous works (Caridakis, et al., 2008) (Castellano & Mancini, 2009) also used Hartmann *et al.* (2005) expressivity parameters. In those works, expressivity is regarded as emotional parameters even though those parameters are based on Gallaher (1992) work. Instead, we aim at capturing expressivity of a person which involves style and emotion. We are estimate expressivity parameters for the whole motion rather than individual gesture in the motion.

Chapter 3 - Estimation of Expressivity Parameters

In this section, we describe our approach to estimating expressivity of a person. Hartman *et al* (2005)expressivity parameters are used to create synthesized motion. We are estimating three of those parameters from real human motion. Overall our estimation process is a three stage process. They are as follows

a) Learning the expressivity estimation from the synthesis motion
b) Estimating expressivity from a real human motion and animating an ECA
c) Testing our results with user survey

In this chapter we are going to discuss about the first step, learning the expressivity estimation from synthesis motion.

3.1 Learning Expressivity Estimation

Before estimating from a real human we learn the expressivity from synthesized motion. Learning process will tell us accuracy of the estimation method. Figure 3-1 shows the step by step learning process.

As a first step, we create the reference corpus using Greta Animation Engine with known expressivity parameters as input. In the process of creating synthesis motion, we also have joint angles as output from the animation engine. Using forward kinematics we convert the joint angles to wrist positions. We design our expressivity estimator which process the input wrist position and delivers expressivity parameters as output. Then, we compare the estimated expressivity parameter with input expressivity parameter and error is measured in terms of absolute mean error. If the estimated expressivity parameters not validated with input expressivity parameters then we need to improve the estimator. If estimated parameters are validated then our estimator is ready we can proceed to the next stage.

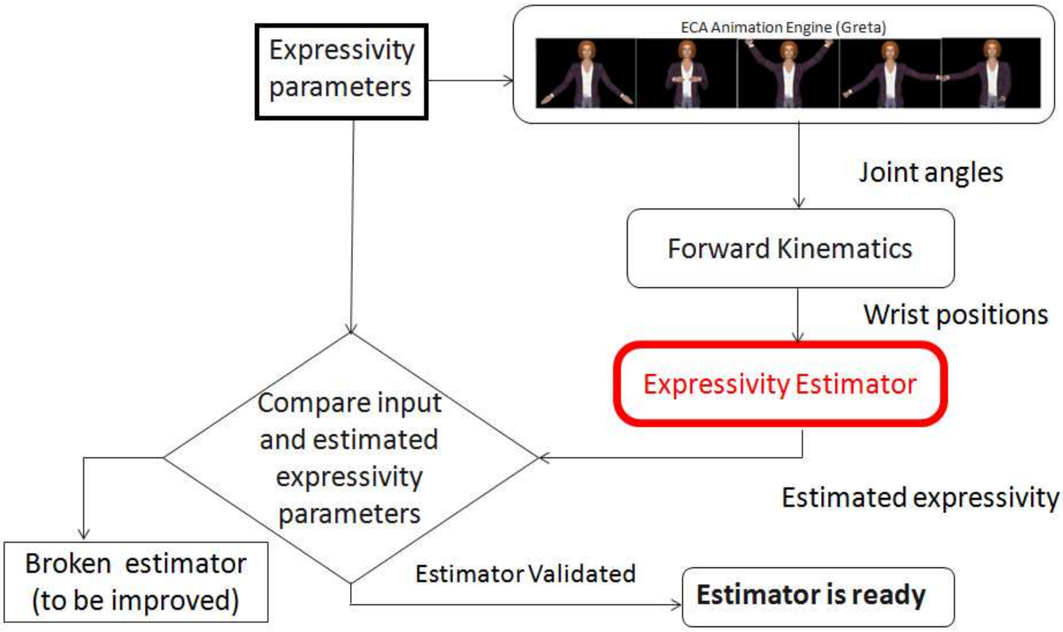


Figure 3-1 Step by Step learning process

3.1.1 Reference Corpus

The purpose of this reference corpus is to learn how to estimate the expressivity parameters. To create the reference corpus we use Greta animation engine. The input for Greta animation engine is through Greta editors and players.

3.1.1.1 Greta Editors and Player

Greta gesture editor creates the individual gestures. Screen shot of the gesture editor is shown in Figure A-1. A gesture can be constructed by three phases namely preparation phase, stroke phase and retraction phase. Each phase can be characterized by the hand shape, hand orientation, arm position and hand orientation. Also we can define the gesture for left arm or right arm or both arms. Along with this expressivity parameters are given as input. The output of the gesture editor will be visualization of the gesture and upper body joint angles for the gesture. We created nine individual gestures with varying spatial and temporal extent. Spatial and temporal extents are varied in the range from -1.0 to +1.0 in the interval of 0.25.

These individual gestures are combined using BML editor to create a full motion. Screen shots of the BML editor is shown in Figure A-2. The created full motion is can be visualized in Greta player the screen shot of the Greta player is shown in Figure A-3. Once the motion is created we have joint angles. Using forward kinematics wrist positions are calculated from joint angles. The detailed explanation of the reference corpuses are given in section 3.4.1.1, section 3.4.2.1 and section 3.4.3

3.2 Expressivity Estimator

Once we have the wrist positions of the hands, we are ready to estimate the Hartmann *et al*. (2005) expressivity parameters. There are six parameters out of which we estimate and implemented three parameters videlicet spatial extent, temporal extent and repetition to describe the hand motions of the person.

Spatial extent defines the amplitude of movements like expanding or contracting the space for hand motions. It is a measure of the amplitude of the movements which is usage of 3D space. Equation given in section 2.5.2.1 is to find the new wrist positions with spatial extent as input. In our case, we know the wrist positions we determine the unknown spatial extent by back substituting the known wrist positions which is explained in detail in section 3.4.1.

Temporal extent talks about duration of movements either it is taking more time to execute the motion or less time to complete the motion. This is measured across successive poses in the motions. Temporal extent is estimated from instant speed. Instant speed is calculated between successive poses. We need to measure the temporal extent from the meaningful phase of the gesture. That is during the stroke phase. Since we are not segmenting the motions to individual gestures, we are going to determine the temporal extent from the instant speed of the whole motion. The detailed methodology we used to determine from temporal extent from instant speed is given in section 3.4.2

Repetition parameter tells number of times a stroke being repeated while making a gesture. The detailed explanation of estimation process is explained in the section 3.4.3

3.3 Validation Process

This estimation is method is validated against synthesis motion. We describe how the gestures are generated through Greta animation engine.

A gesture is generated for key positions. A key position is comprised of shoulder, elbow and wrist joint angle positions. A key position is denoted in terms of its wrist positions. For a gesture to be generated a set of key positions are chosen. These key positions are interpolated through spline. The generated gesture has neutral expressivity parameter value. Varying on expressivity parameter values, the gestures are generated.

3.3.1.1 Expressivity Parameter Validation

This estimated method is validated against synthesis motion generated using Greta (Hartmann, Mancini, & Pelachaud, 2002) with input number of repetition, spatial and temporal extent. As explained before the gestures are edited using the BML editor and their animation is computed. We have created a test corpus to validate our estimation process. From the generated motions we find the wrist positions through 3D motion capture algorithm (Gómez Jáuregui, Horain, Rajagopal, & Karri, 2010). We estimated repetition, spatial and temporal extent using our method. Since we know the input number of repetitions, input spatial and temporal extent values, we compare our estimated repetition, spatial and temporal extent with it and error is measured.

We will give detailed explanation of estimation and validation process in the upcoming sections.

3.4 Estimation of Expressivity Parameters

We propose a new algorithm to estimate spatial extent, temporal extent and repetition parameter from the real human. We estimate these three expressivity parameters from 3D hand motion trajectories. We use the Greta animation engine to generate example motions with controlled expressivity, through which we our algorithms are validated.

3.4.1 Estimating Spatial Extent

Hartmann *et al.* (2005) defines expressivity parameters as one of the input to the synthesized animation. These parameters deliver the identity and mood of a person. The spatial extent defines a scale of coordinates to the wrists positions with respect to an origin. Following the convention of the Greta animation engine, we set that origin at the sacroiliac vertebra, which stands approximately between the rest positions of the two wrists. It is estimated from hand 3D motion trajectories.

The definition of spatial extent and its mathematical explanation is given in section 2.5.2.1 rearranging equation (2.1) after taking mean we obtain (3.1) where the term *SPC* is to be estimated:

$$(1 + SPC \cdot Agent_{\blacksquare}) = \frac{\overline{p'_{\blacksquare}}}{\overline{p_{\blacksquare}}} \quad \textbf{(3.1)}$$

here ■ represents the lateral, vertical and frontal directions. where $\overline{p_{\blacksquare}}$ can be learnt from a set of motion trajectories generated with null spatial extent. Finally, we estimate the spatial extent *SPC* of our input trajectory as the mean of the directional spatial extents:

$$SPC = \frac{1}{Agent_{\blacksquare}} \left[\frac{\overline{p'_{\blacksquare}}}{\overline{p_{\blacksquare}}} - 1 \right] \quad \textbf{(3.2)}$$

Finally, we estimate the spatial extent *SPC* of our input trajectory as the mean of the directional spatial extents:

$$SPC = \frac{1}{3}(SPC_x + SPC_y + SPC_z) \quad \textbf{(3.3)}$$

3.4.1.1 Validating Spatial Extent Estimation

We evaluate the above approach for estimating the spatial extent against a corpus of 81 communicative hand motions with controlled expressivity that are generated using the Greta animation engine (Hartmann, Mancini, & Pelachaud, 2002) . As a first step in generating communicative hand motion, we generated nine different gestures using Greta editor (Figure A-1 and). These nine gestures are generated with same expressivity value (for example say, Spatial Extent = 0.0). We combined these nine gestures through BML editor (Figure A-2) and made it as single motion.

Motion trajectories are generated with nine different values of spatial extent (-1.0, -0.25,-0.75, -0.5, -0.25, 0.0, 0.25, 0.5, 0.78, 1.0) and nine values of temporal extent (-1.0, -0.25,-0.75, -0.5, -0.25, 0.0, 0.25, 0.5, 0.78, 1.0). Thus results in 81 trajectories. Now we have eighty one different motions having the spatial extent values ranging from -1.0 to +1.0 in the interval of 0.25 with respect to temporal extent ranging from -1.0 to +1.0 in the interval of 0.25.

When these motions are animated through BML editor, BML editor gives Body Animation Parameter (BAP) and Face Animation Parameter (FAP) files as output. The upper body joint angles information are stored in BAP file in encoded form. We decode the joint angles from BAP file and applied forward kinematics algorithm (Craig, 1986) to find the wrist positions. From the wrist positions, spatial extent is calculated as described in paragraph 3.4.1. We examine the error difference between the estimated spatial extent and the actual spatial extent which is given as input in gesture editor. The estimated spatial extent and actual spatial extent for synthesized motions are plotted and it is shown in Figure 3-2.

From Figure 3-2 we see that estimated spatial extent not exactly lying with actual spatial extent due to error in the estimation process. The error in the estimation process is calculated in terms of absolute mean error. The absolute mean error for spatial extent estimation is 0.14. In the range [-1.0, +1.0] there are 200 samples of size 0.01. It is understood that our spatial extent estimation process produces the result with 7 % of error.

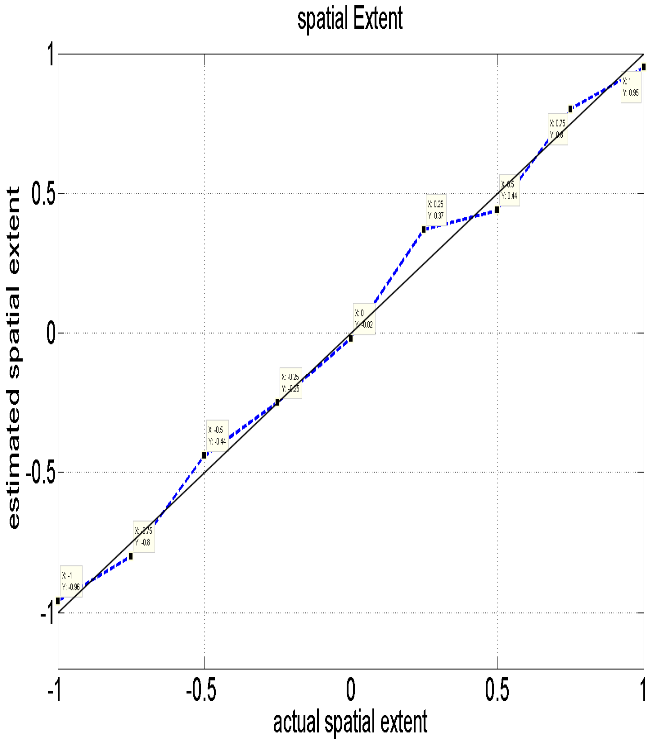


Figure 3-2 Estimated spatial extent vs input spatial extent for synthesized motions

3.4.2 Estimating Temporal Extent

As per the definition of temporal extent in section 2.5.2.2 the higher the temporal extent *TMP*, the higher the speed is. From this tendency, we derive a heuristic to estimate *TMP* from the observed instant speeds along the trajectory, *i.e.* the distance between poses at successive time intervals.

In the work by Hartman’s *et al.*, (2005) expressivity parameters affect synthesized gestures during stroke time only. Rather than segmenting strokes, which known to be a difficult problem (Quek, McNeill, Bryll, Kirbas, & Arslan, 2000), we estimate expressivity parameter from whole motion trajectories. In order to do that we generate a reference corpus containing nine set of synthesized motions through gesture editor and BML editor as explained in the section 3.4.1.1. Nine set of synthesis motions in the reference corpus will have the temporal extent value ranges from -1.0 to +1.0 in the interval of 0.25. Some of the poses in the reference corpuses are shown in Figure 3-3. As we said before, the BML editor which combined the motions will give output as BAP files. After decoding the BAP files we calculate the wrist positions through forward kinematics (Craig, 1986). From synthesis wrist trajectories, we find that high *TMP* values give high speed only during the strokes, while the preparation and retraction exhibit low speeds.

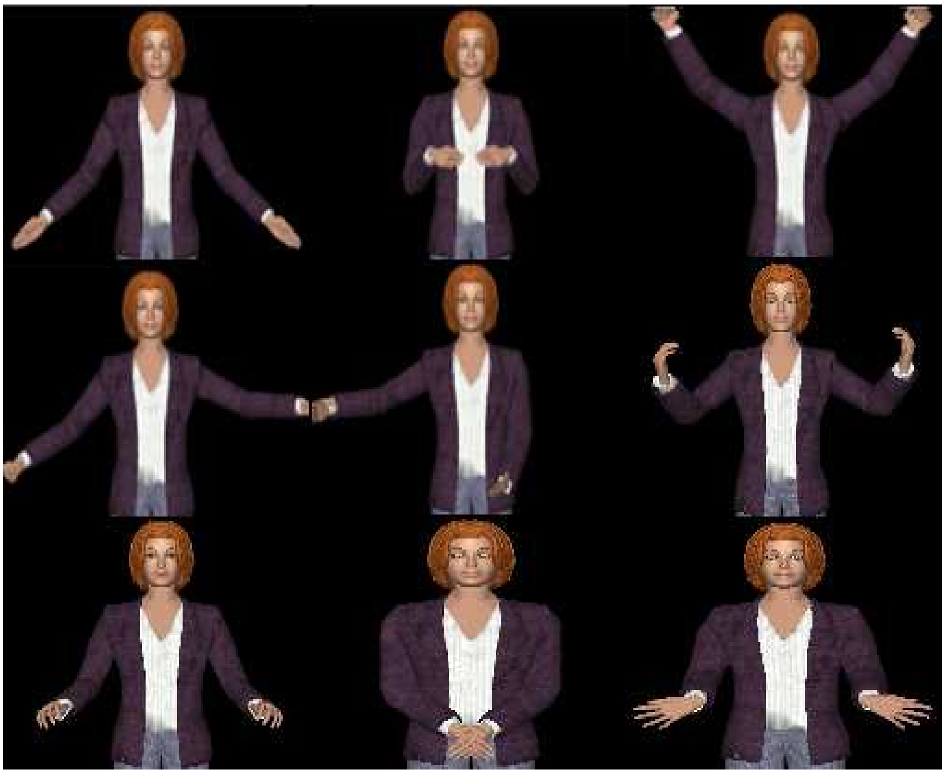


Figure 3-3 some of the poses in reference corpus

Therefore, *TMP* can be estimated by considering only some quantile of the higher speeds. Based on this, we sort the instant speeds of nine different motion trajectories in the reference corpus in descending order. We set the quantile limit based on the correlation between quantile and the input temporal extent value in the gesture editor where reference corpus is generated.

From the learnt trajectories of the reference corpus, temporal extent is estimated from 7 % of upper quantile. The correlation for mean speed up to 7 % upper quantile is 93%. After rejecting 2 % upper quantile, the correlation improves to 97 %. This improvement is because there may be a chance for discontinuity while combining gestures to form a motion in BML editor (Figure A-2). Due to this correlation value is increased. So we reject initial 2 % of upper quantile for the mean calculation and mean speed with the remaining 5% of upper quantile we achieve best mapping to temporal extent in the range [1.0, +1.0] through linear regression. Higher instant speed among reference corpus synthesized motions is shown in Figure 3-4.

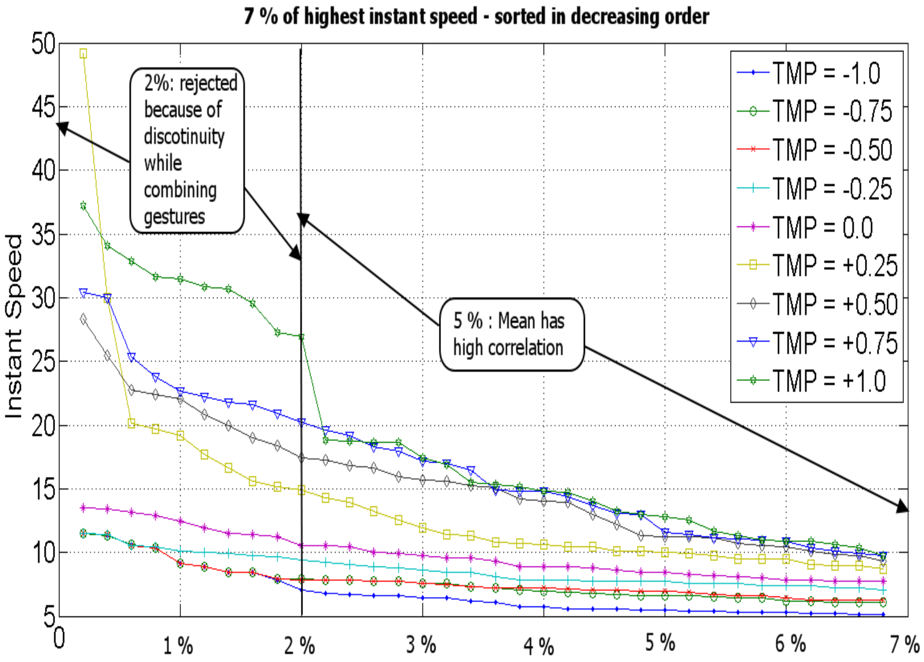


Figure 3-4 Higher instant speeds from synthesized motions

3.4.2.1 Validating Temporal Extent Estimation

The estimation of temporal extent is also validated against synthesized motion with varying controlled expressivity. We generated another set of corpus called testing corpus which contains the nine set of motions in which each motions contains nine gestures. Nine motions have the temporal extent value ranging from -1.0 to +1.0 in the interval of +0.25. Some of the poses in testing corpus are shown in Figure 3-5.

The idea is to estimate the temporal extent from the testing corpus from the heuristics described in section 3.4.2 and to make the comparison between estimated temporal extent and actual temporal extent value. As usual first step is to find the wrist positions from the synthesized motion. As we explained before we determine the wrist positions through forward kinematics.

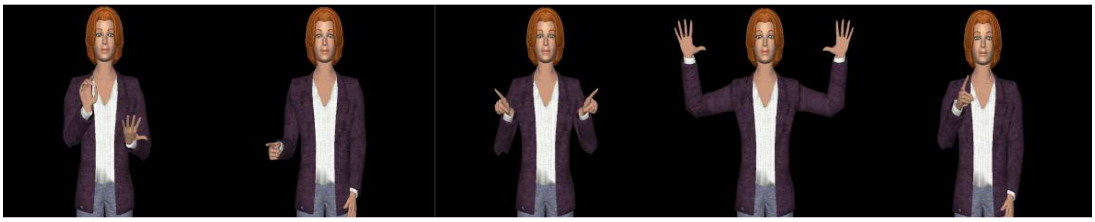


Figure 3-5: Set of poses in Testing Corpus

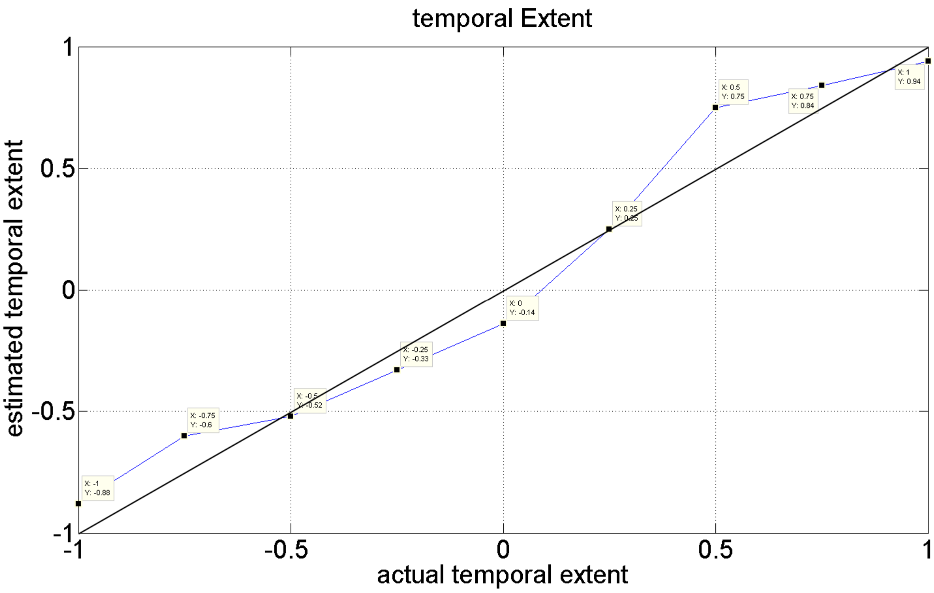


Figure 3-6: Estimated temporal extent vs input temporal extent for synthesized gestures

From the wrist positions we calculate the instant speeds of the wrists among each pose. We map instant speed to temporal extent values using the linear regression as described in section 3.4.2. The estimated and actual temporal extent values are plotted in Figure 3-6. Similar to the spatial extent, the error measure for *TMP* is also done by absolute mean error. Absolute mean error for *TMP* estimation is 0.15. (i.e.,) our *TMP* estimation method causes 7.5 % error in determining the TMP from 3D motion data.

3.4.3 Estimating Repetitions

The repetition parameter is to be estimated from wrist 3D trajectories.

When a gesture is being repeated, only its stroke is repeated, not the preparation or the retraction phases (Efron, 1941). During repetition, the wrist trajectory follows or is close to the path at the earlier stroke, with some delay.

One way of finding the repetition in the motion trajectory is through Fourier transform. If the repetition occurs then there will be a periodicity in the stroke period. The windowed Fourier transform allows detecting periodicity that lasts for a time interval. Unfortunately, the duration of repetitions varies. Furthermore, the number of repetitions is very small (2 or 3 periods only). Finally, we see from Figure 3-7 that, in case of 2 repetitions, the second repetitions completes faster than the first repetition. The time period varies among repetitions. Because of these three reasons, windowed Fourier transform fails in finding the repetitions.

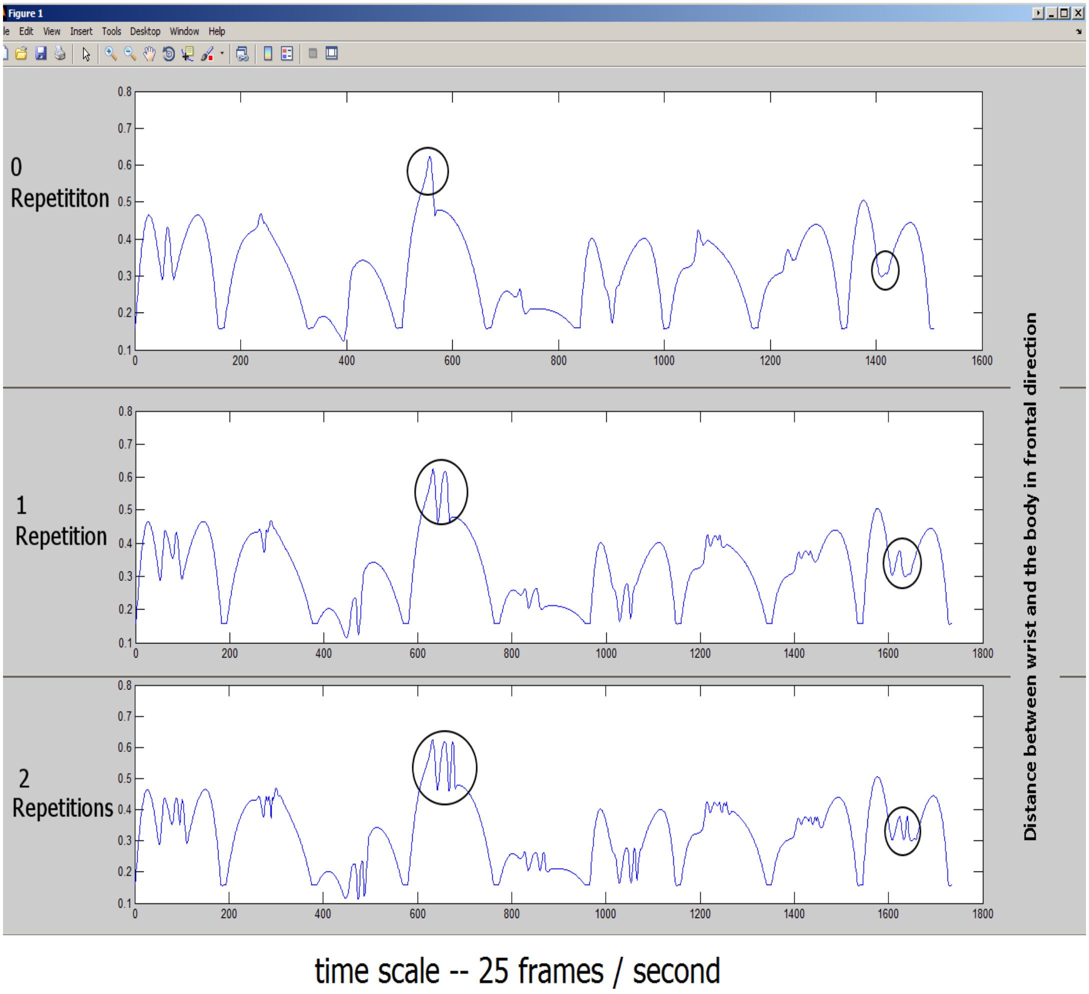


Figure 3-7: Stroke repetitions in wrist trajectory. Here, only the frontal coordinate of the wrist 3D trajectory is plotted. Some of the repetition phases are circled.

Another way of estimating repetition is through windowed auto-correlation. Correlation is a measure of the similarity between two or more variables. The correlation coefficient ranges from -1.0 to +1.0. Higher absolute value of the correlation indicates that the variables are in a linear relationship. In our case, if repetitions happens, then correlation will be high.

For windowed auto-correlation, we compute correlations over a sliding time interval. The window length should match the stroke length. Unfortunately, as we already noticed, in case of two repetitions strokes duration varies (Figure 3-7). Even if we choose correct window length we may fail in identifying the 2nd repetition.

Figure 3-8 and Figure 3-9 show some experimental results of correlation for single and double repetition, respectively. Since we are interested only in positive correlation, negative correlations are clipped. Window length is one second. Double repetition has less positive correlation than single repetition. Correlation fails to find the second repetition because time difference between stroke and 1st repetition differs from time difference between 1st repetition and 2nd repetition. We couldn't find number of repetitions in this method, so correlation won't suitable for irregular time period among repetitions.

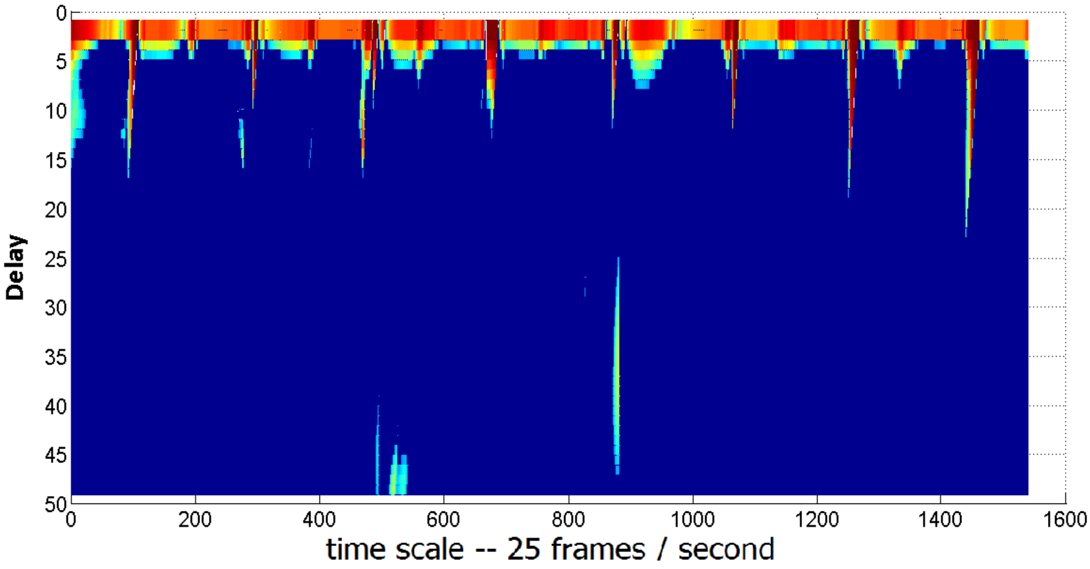


Figure 3-8: Example correlation with one repetition. Correlation over a sliding window between the current wrist 3D position at the current frame (given in abscise) and the upcoming wrist positions (delay given in ordinate). The wrist trajectory is sampled at 25 positions per second.

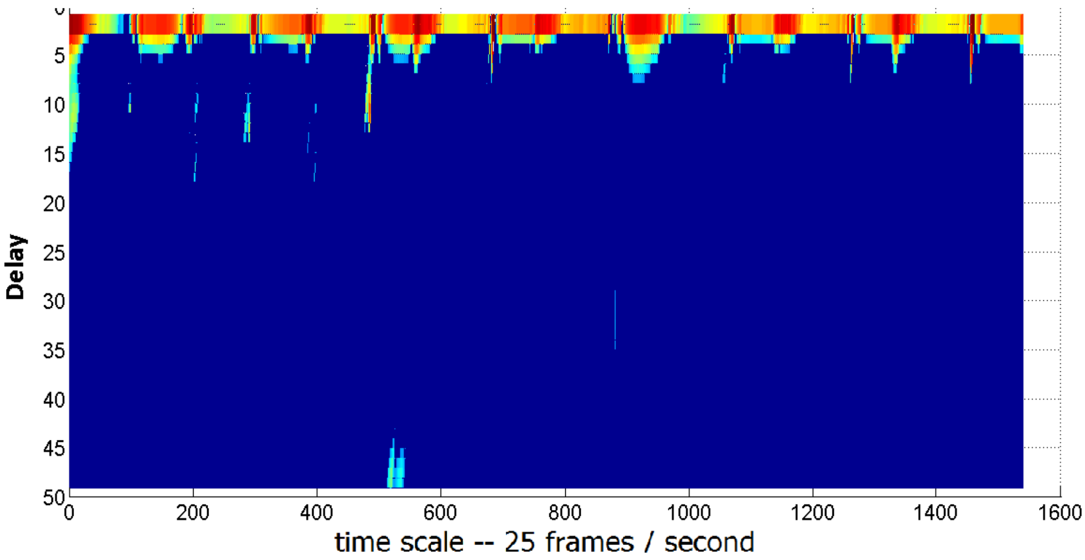


Figure 3-9: Example correlation with two repetitions. Correlation over a sliding window between the current wrist 3D position at the current frame (given in abscise) and the upcoming wrist positions (delay given in ordinate). The wrist trajectory is sampled at 25 positions per second.

If a stroke is repeated, the distance between the wrist positions in the current and the repeated strokes will be small. Repetitions can then be detected by comparing the current and upcoming trajectory over a sliding time window, the duration of which is a bit shorter than the stroke duration. The maximum distance over the sliding window could be used to detect repetition, although it would be highly sensitive to isolated noisy poses. We rather use the mean distance over the sliding window. Since distances are positive, their sum can be small only if each distance is small.

Note that no segmentation of the wrist trajectory into stroke is involved here. In addition, note that rest periods, where the hand keeps still, also lead to zero distance. Although, rest periods can be distinguished from repetition since they have a longer duration than the repeated strokes.

Just like for the previous expressivity parameters, we generate a reference corpus of wrist 3D trajectories using the Greta animation engine. This corpus consists of three hand motions, each composed of nine gestures, having repetition value of 0, 1 and 2. Figure 3-10 shows some of the poses in this corpus.

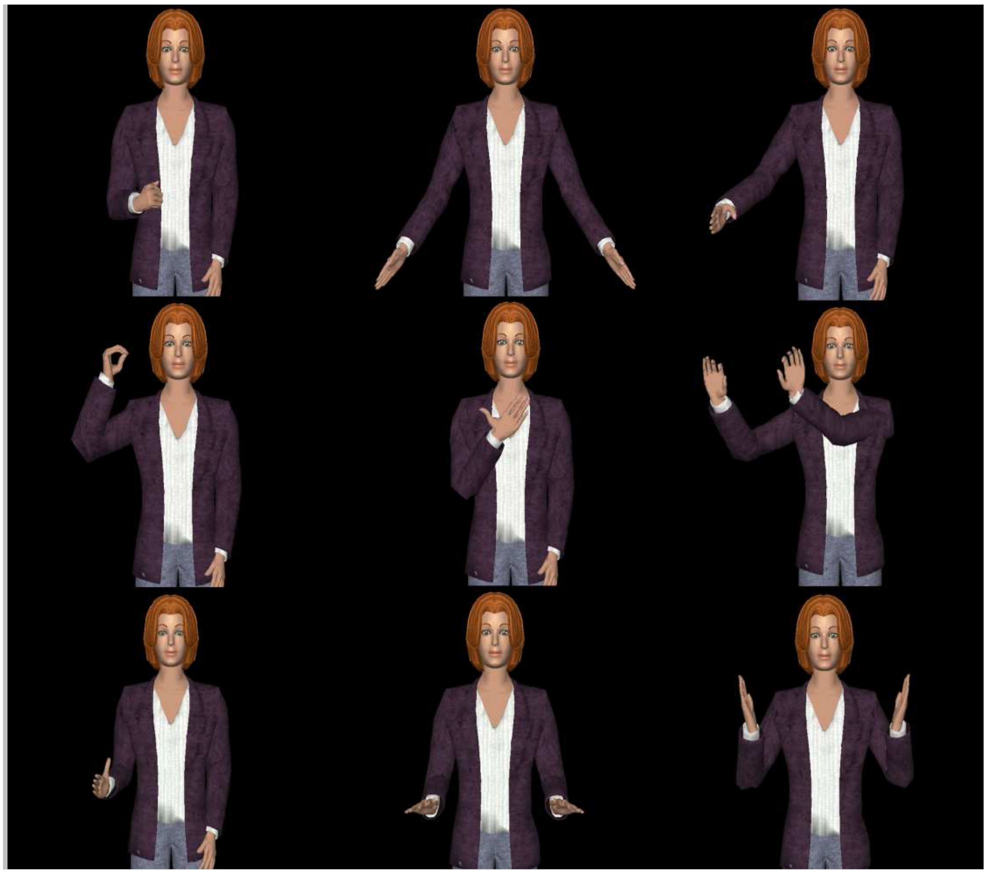


Figure 3-10: Set of poses in Reference Corpus for Repetitions (available online from www-public.it-sudparis.eu/~rajagopa/RepVideos.htm).

Repetition phases can be seen in Figure 3-7 that shows the frontal coordinate of wrist 3D trajectories in the corpus. Some repetition phases have been circled. Our aim is to detect such phases and count the number of repetitions in them.

From the reference corpus, we find the duration of a (possibly repeated) stroke is around half a second (Figure 3-7), so we use a slightly shorter sliding window of 10 frames (at 25 poses per second). The resulting averaged distances are shown as functions of the current frame number and the delay in Figure 3-11.

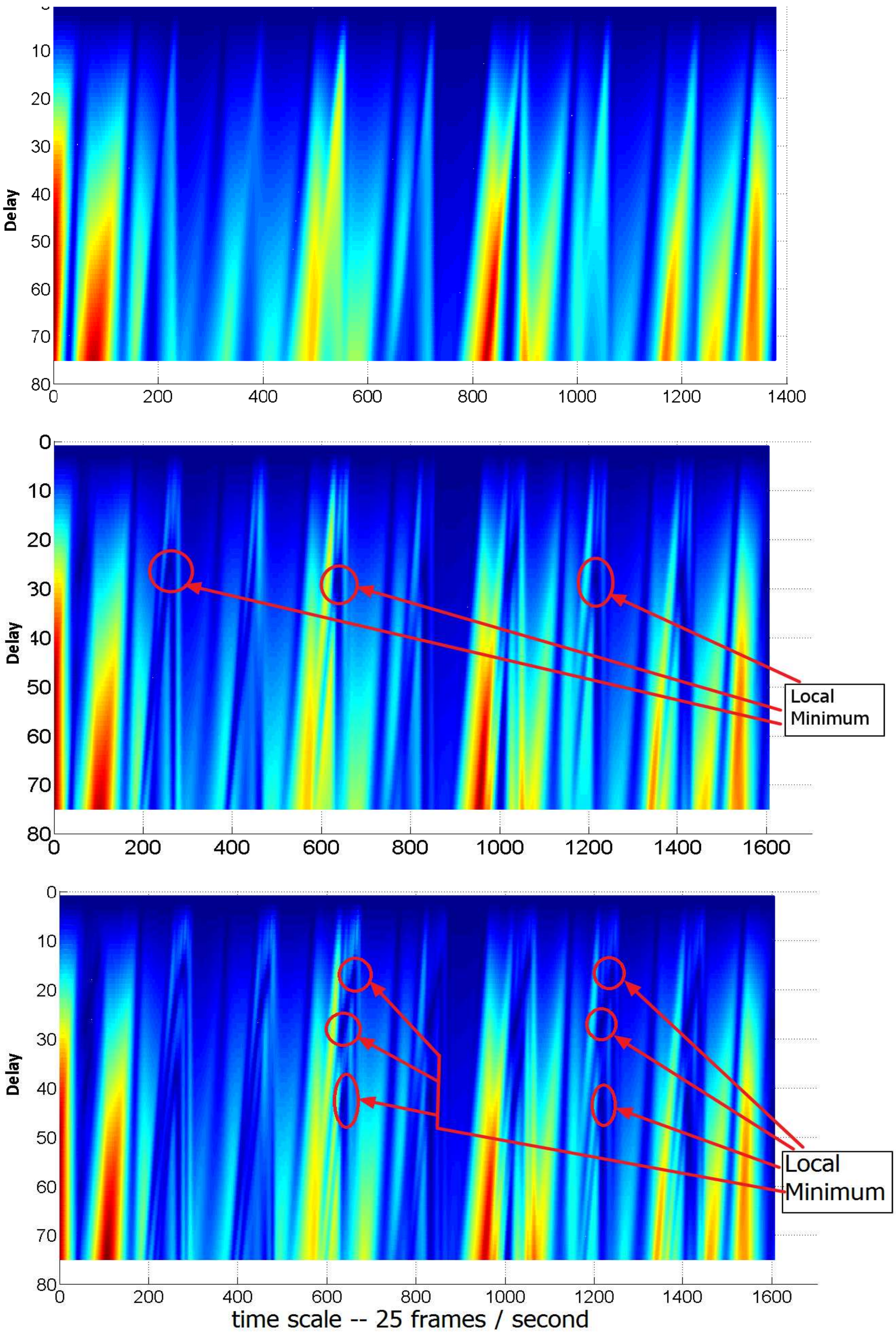


Figure 3-11: Example distance with zero, one and two repetitions, respectively. The average over a sliding window of the distances between the current wrist 3D position at the current frame (given in abscise) and the upcoming wrist positions (delay given in ordinate). The wrist trajectory is sampled at 25 positions per second.

Figure 3-12 gives a schematic view of the distance configuration for repetitions. The Euclidean distance is found by moving the sliding window. When the sliding window reaches the repetition phase, the Euclidean distance will be less. Those regions are shown in white patches in the graph. Repetitions appear as short interruptions with low values in the oblique and vertical lines of high values. Our algorithm should be capable detecting these interruptions in the high values.

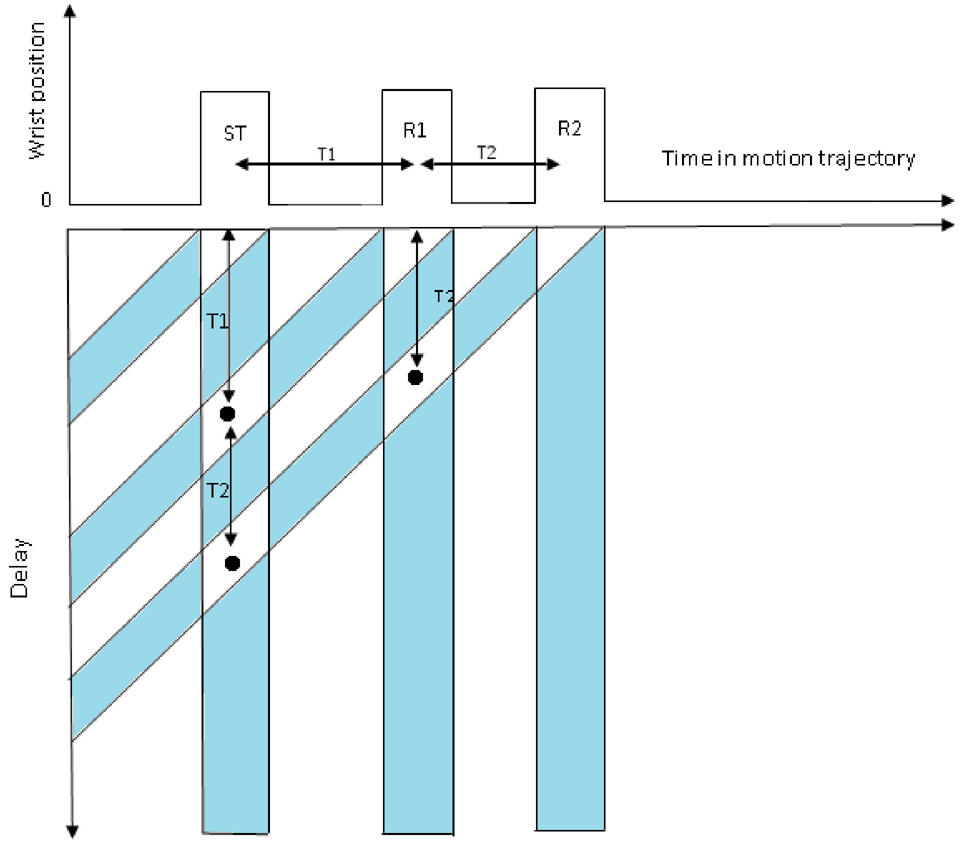


Figure 3-12: Schema of wrist distances in case of repetitions.
Upper part: wrist trajectory; ST is the initial stroke, while R1 and R2 are respectively repetitions after T1 and T2 time shifts, respectively.
Lower part: distances between wrist positions with respect to time and delay.
Higher distances are shown in blue, while lower distances are in white.

Repeated strokes appear as distance local minimums close to zero in the Figure 3-11. These could be detected by thresholding. But, rest periods, where no wrist movement occurs, would be selected along with repetitions. We need to discard rest periods.

From Figure 3-7, repetitions appear as low values surrounded with high values. We apply mathematical morphological technique to locate those minimum Euclidean distances. Also we should avoid the minimum Euclidean distances due to rest positions. Such confusions can be avoided using top hat transform and repetitions are detected.

In image processing, top hat transform is useful for enhancing the detail in the presence of shading (Gonzalez & Woods, 1992). This transformation is also good in 1D for finding peaks that are say, greater than a certain width and more than a certain depth (significant peaks). Top hat transform identifies the local minimums. In our case, during the repetition the distance between the consecutive poses will be minimum compared to the non repetition period in the hand motion. Top hat transform detect the minimum distance values present in between maximum distance values. .So, top hat transform is employed to detect the repetition in the wrist 3D trajectory.

From Figure 3-12, the repetitions are identified when vertical column and obliged column cross each other. We perform top hat transform on these distances in two ways. In one way, we move the structural element of the top hat transform is vertically moved along the wrist trajectory and in the second way we move the structural element diagonally in the wrist trajectory. Multiplying these two will result detection of repetitions.

For double repetition, when we move the structural element of the top hat transform starting from stroke period, after 'T1' time, first repetition is detected as peak. After 'T2' time, second repetition is detected as another peak. 'T1' is time between stroke 'ST' and first repetition 'R1'. 'T2' is the time between 'R1' and second repetition 'R2' in Figure 3-12. When we move the structural element starting from first repetition, we will detect another peak after time 'T2. This is due to second repetition. If there is a double repetition, then we will have three peaks after performing top hat transform.

We experiment top hat transform in vertical column and obliged column with the synthesis motion, the results are shown in the Figure 3-13, Figure 3-14, Figure 3-15 and Figure 3-16. Along with the peaks we have the vertical lines in both the vertical column and obliged column transformation. These lines are due to local minimums present in the Figure 3-11. When we multiply these two we will end up with the blobs which show repetition alone.

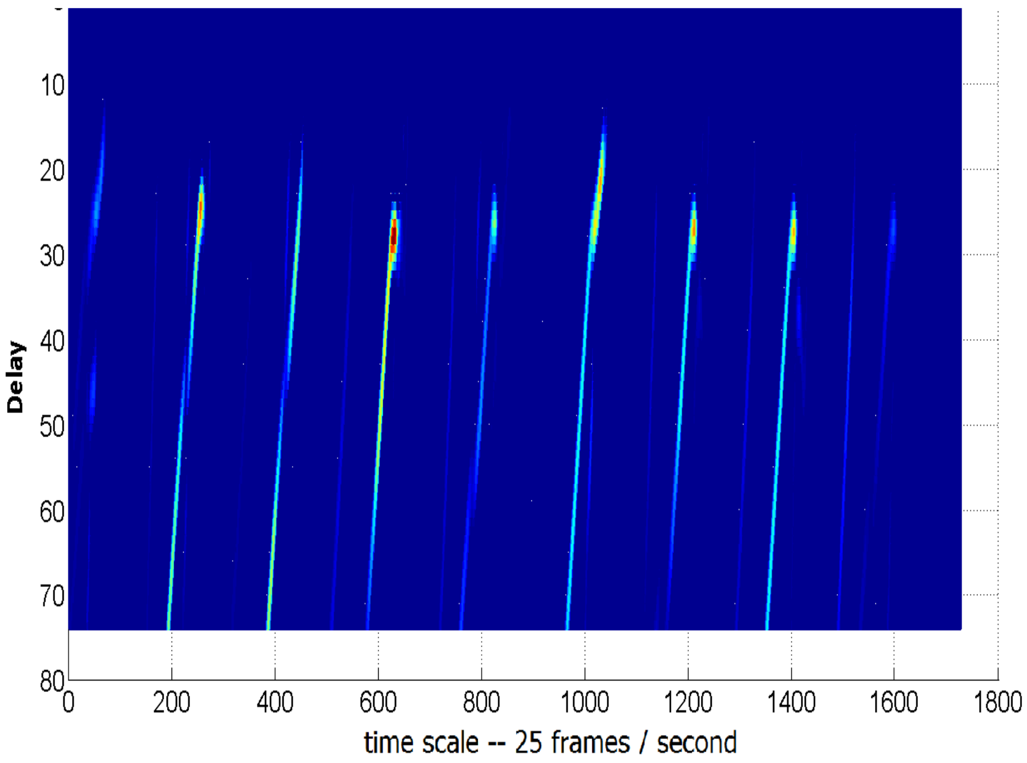


Figure 3-13 Top Hat transform – structural element along vertically for 1 repetition

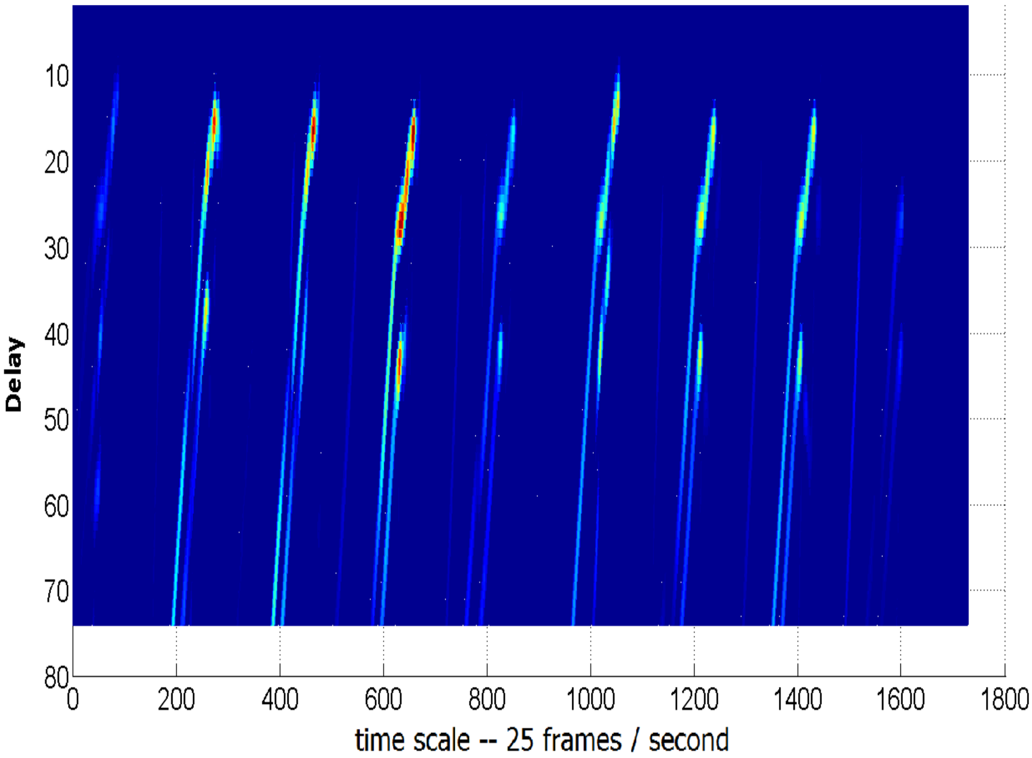


Figure 3-14 Top Hat transform – structural element along vertically for 2 repetitions

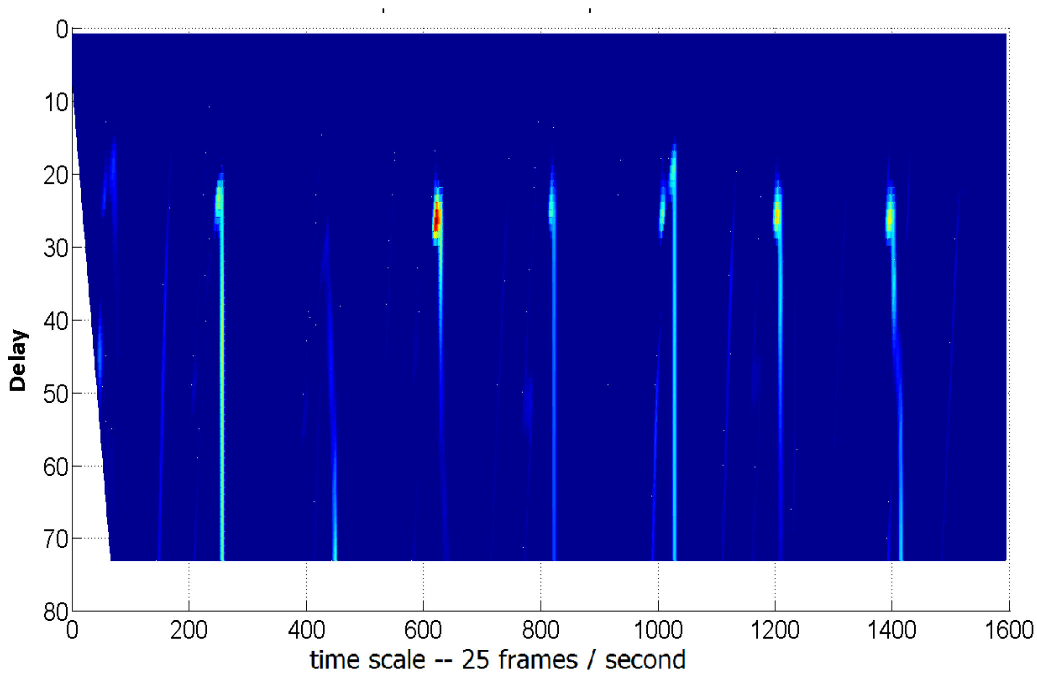


Figure 3-15 Top Hat transform – structural element along obliged for 1 repetition

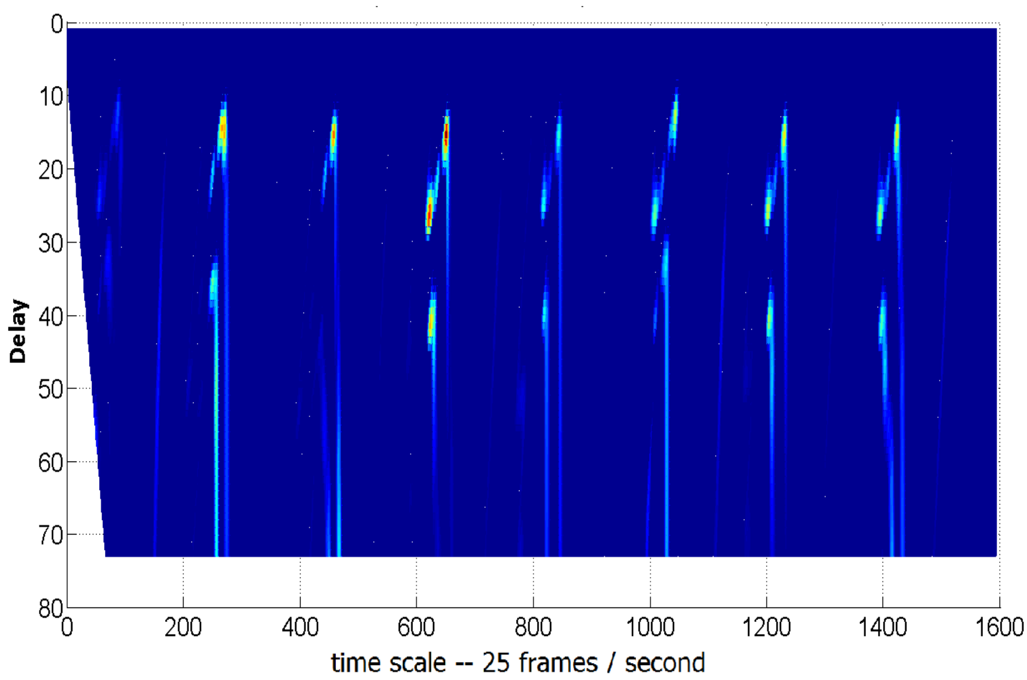


Figure 3-16 Top Hat transform – structural element along obliged for 2 repetition

We multiply the both (structural element vertical and obliged) the top hat transforms to highlight the peaks. The resultant of that is shown in Figure 3-17 and Figure 3-18. From these graphs, the repetitions are shown as blobs in the wrist trajectory. Also we can count exactly the number of repetitions present in the trajectory.

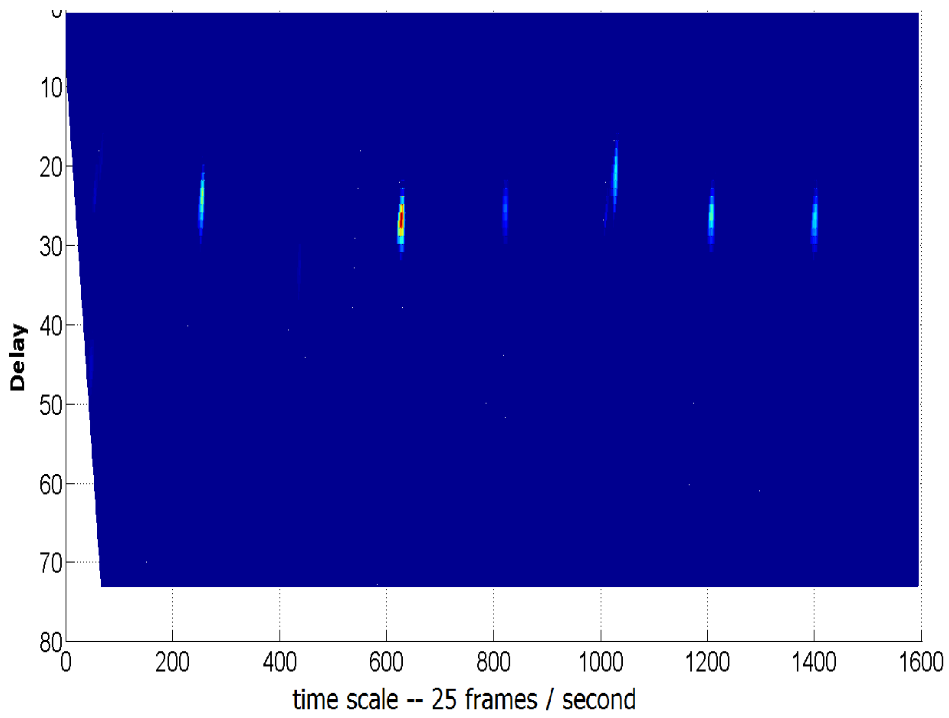


Figure 3-17 Top Hat transform – vertical X diagonal structural element for 1 repetition

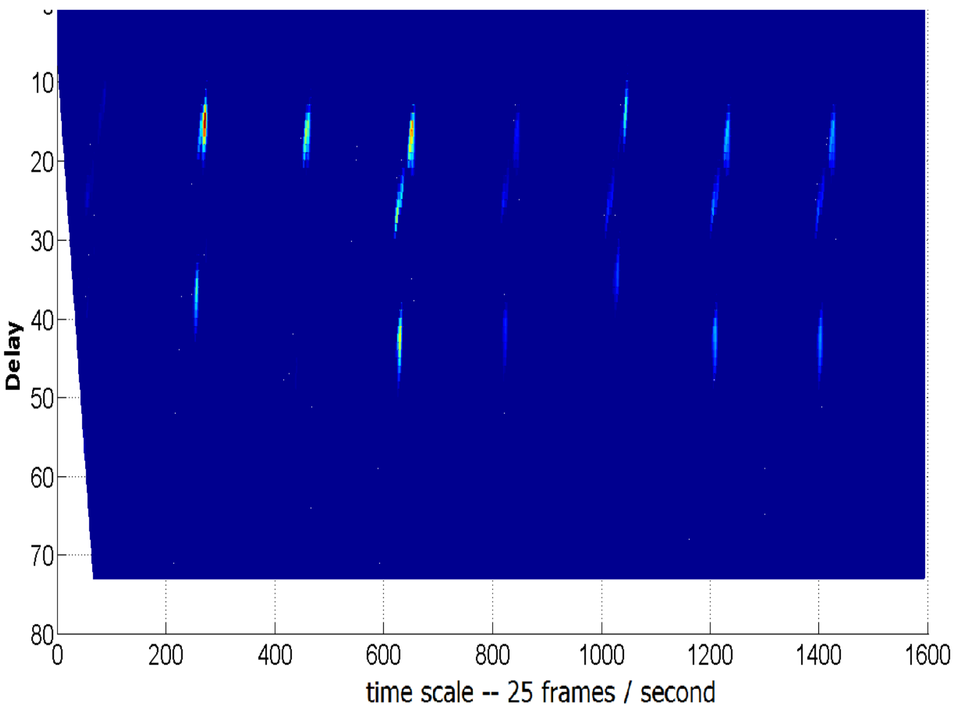


Figure 3-18 Top Hat transform – vertical X diagonal structural element for 2 repetition

Since the estimation process yields fixed values of repetition (0, 1 or 2), the validating estimation process is not required. We tested our estimation process directly on real human motion trajectory

3.5 Conclusion

In this chapter, we estimate three expressivity parameters, namely spatial extent, temporal extent and repetition, from wrist 3D trajectories. For estimating the spatial and temporal extents, we generated the four different hand motion corpora. Out of four different corpora two of them are reference corpora and the remaining two are test corpora. These corpora contain communicative gestures. In the first reference corpora, spatial and temporal extents are varied as explained in the section 3.4.1.1. In the second reference corpora repetition is varied as 0, 1 and 2. Our estimation processes for three parameters are explained in sections 3.4.1, 3.4.2 and 3.4.3.We validated our estimation process with test corpora.

.

Chapter 4 - Perceptive study

In chapter 3, we have presented a computational approach to estimate expressivity from motion trajectories. To assess the importance of expressivity for cloning, we need to evaluate the reverse perception, that is whether users can recognize persons based on their gesture expressivity.

In this chapter, we experimentally study user perception of gesture expressivity. Gesture expressivity of real humans is derived from 3D motion capture, and then fed to an avatar animation engine. Users are asked to match synthesis animations that differ only by gesture expressivity with videos of real humans.

4.1 Experimental setup

Hereafter, we first describe the 3D motion capture we used. Wrist 3D positions are then input to the expressivity estimator and, the expressivity parameters are fed as input to the Greta engine to animate a virtual agent. As a result, we have a virtual avatar animated with the expressivity captured from real humans. This work flow is shown in Figure 4-1. Finally, the generated synthesized motion is presented to users to experimentally evaluate this gesture expressivity as a clue for cloning.

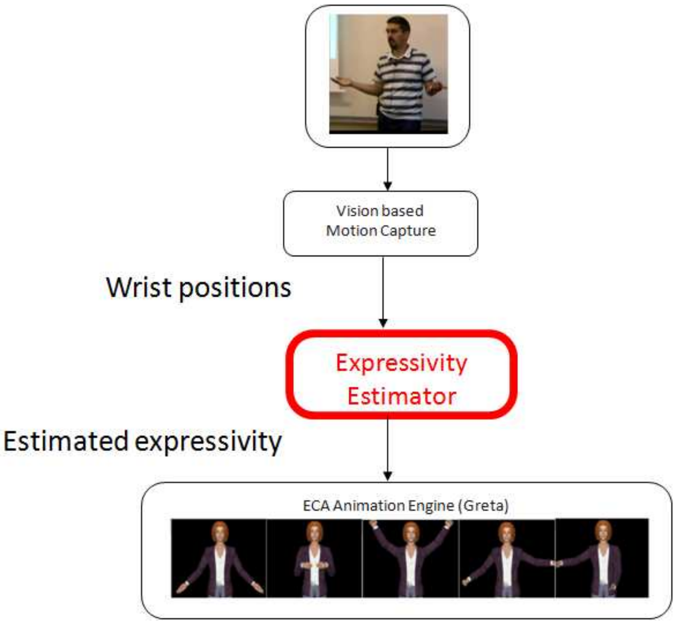


Figure 4-1: Cloning the gesture expressivity of a real human

4.1.1 Capturing Human Motion

To capture the expressivity of real humans, we need to capture their hand motion.

Motion capture was first achieved in the late 1970's for military applications (tracking the movements of pilots) (Furniss, 2000). Optical (Optitrack, 2010), mechanical (Gypsy7, 2010), magnetic (Advanced Motion Measurement, 2010) (Advanced Motion Measurement, 2010), acoustic (Intersense, 2010), inertial (Animazoo, 2010) (Animazoo, 2010) and computer-vision-based systems (Gómez Jáuregui, Horain, Rajagopal, & Karri, 2010) have been designed for that purpose. Since the 1990's, advances in computing power and algorithms have made real-time motion capture possible. Nowadays, motion capture is rapidly becoming cheaper and many more systems have emerged in the market.

Computer vision is attractive for motion capture because it frees the user from any invasive hardware attached to the body. It may work with or without markers, using several cameras or only one and it can work outdoors as well (Moeslund, Hilton, & Krüger, 2006) (Poppe, 2007). Monocular vision has been used for tracking specific motions such as walking, golf swinging, jumping after learning (Agarwal & Triggs, 2006), (Urtasun, Fleet, & Fua, 2006).

We are interested in 3D upper body motion. Recently, Microsoft released Kinect sensors to track body motion (Microsoft, 2010) (Microsoft, 2010). We successfully experimented with a Kinect sensor to track the human motion in live experiments, as shown in Figure 4-2.

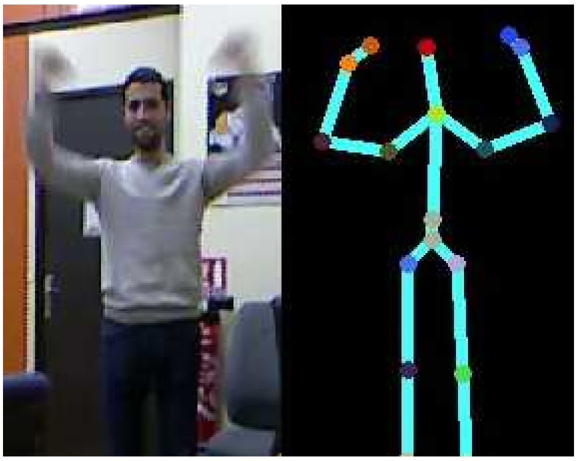


Figure 4-2 Left: Real human motion; Right: Kinect tracking

At the time we started our experiments, Kinect sensors were not yet available. So, we considered the monocular vision system previously developed in our laboratory by Gómez Jáuregui. It allows capturing 3D human motion in real time without any marker from a single video. It proceeds by registering a 3D articulated model of the human body on video sequence, and it outputs wrist 3D positions (Gómez Jáuregui, Horain, Rajagopal, & Karri, 2010).

4.1.2 Estimating Expressivity Parameters of a Real Human

For experiments we considered four videos (V1, V2, V3 and V4) of users performing communicative gestures in front of the camera. Set of poses in those videos are shown in Figure 4-3, Figure 4-4, Figure 4-5 and Figure 4-6. V1 and V2 video clips are parts of video lectures and are about one minute long. The other two videos (V3 and V4) are about 20 second long. These four videos are uploaded here[3].

[3] http://www-public.it-sudparis.eu/~rajagopa/realHumanVideos.htm

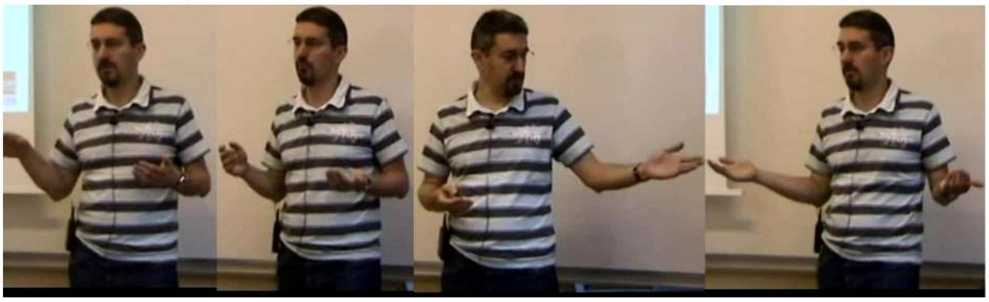


Figure 4-3 Set of poses of real human motion in video sequence V1

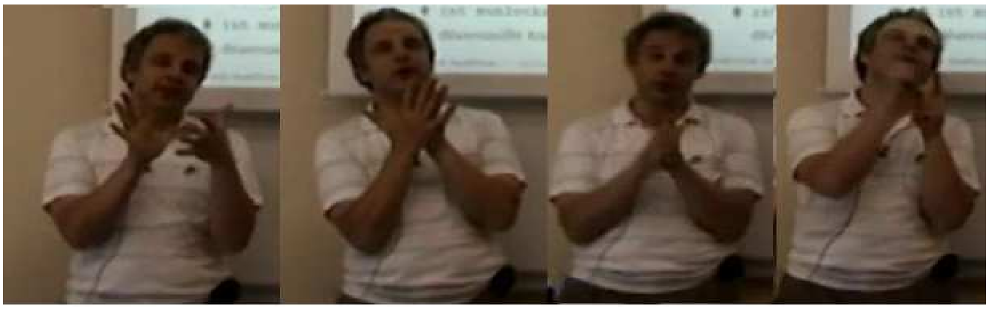


Figure 4-4 Set of poses of real human motion in video sequence V2

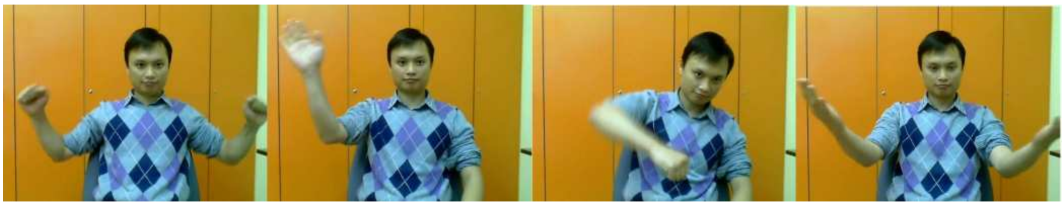


Figure 4-5 Set of poses of real human motion in video sequence V3

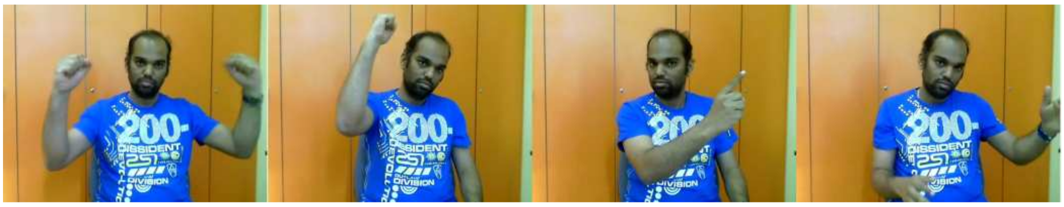


Figure 4-6 Set of poses of real human motion in video sequence V4

The Gómez Jáuregui *et al*. (2010) algorithm tracking algorithm outputs upper body joint angles of the person. A sample result is shown in Figure 4-7. Forward kinematics algorithm (Craig, 1986) then allows to to determine wrist postions from the upper body joint angles.

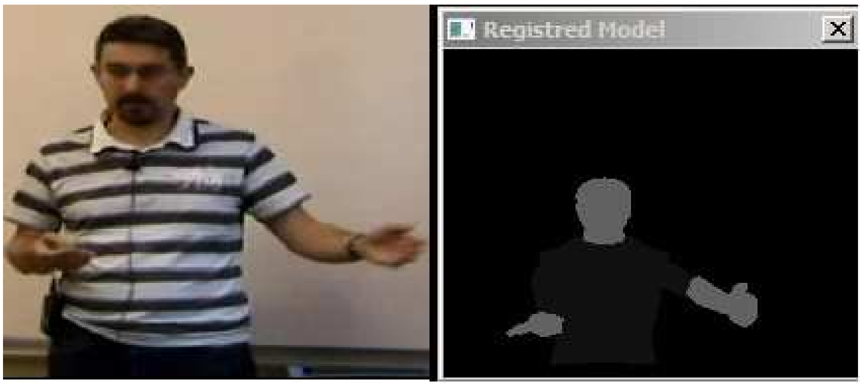


Figure 4-7 Tracking with Gómez Jáuregui *et al.* algorithm

The spatial extent is then estimated as explained in section 3.4.1 for the videos of V1 and V2. As illustrated in the V1 and V2 snap shots, V1 has more spatial extent than V2. The estimated spatial extent for V1 is +0.8 and the spatial extent for V2 is +0.6.

Similarly, the temporal extent is estimated for V1 and V2 as described in section 3.4.2. The estimated TMP is -1.0 for V1 and -0.7 for V2. This shows the user in V2 has faster gestures than the user in V1.

The repetition parameter is estimated as explained in the section 3.4.3 for all the four videos. V1 and V2 do not show any repetitive gestures. Our algorithm yields the same result. Repetitive gestures in V3 have two repetitions, while in V4 they have only one repetition.

4.1.3 Animating the Virtual Human with Expressivity

We used the Greta animation engine (Hartmann, Mancini, & Pelachaud, 2002) to synthesize a communication gesture and to vary the motion based on input expressivity.

We first input to the Greta animation engine the spatial and temporal extent values estimated from videos V1 and V2. This is shown in Figure 4-8 and Figure 4-9, where high and low spatial extents can be seen.

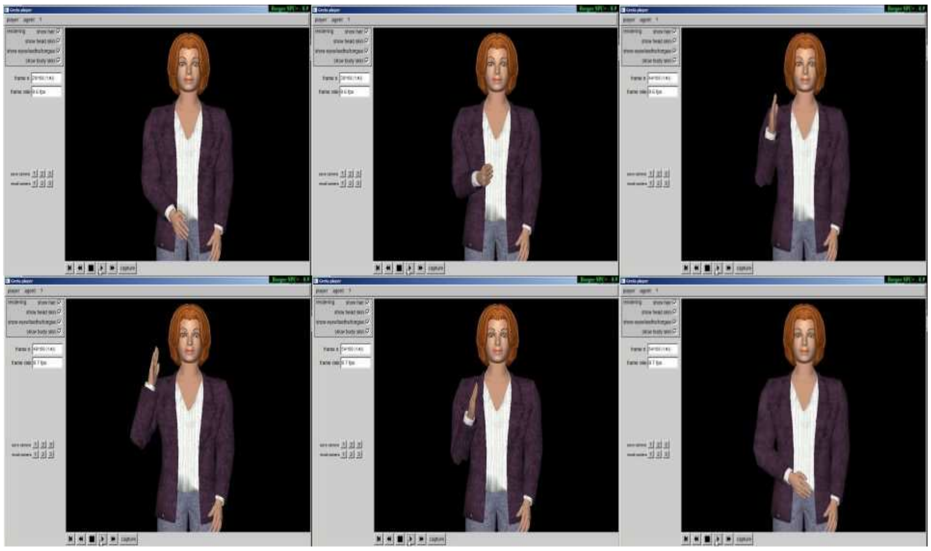


Figure 4-8 Animated gesture using Greta for the spatial extent of V1. Note the large spatial extent.

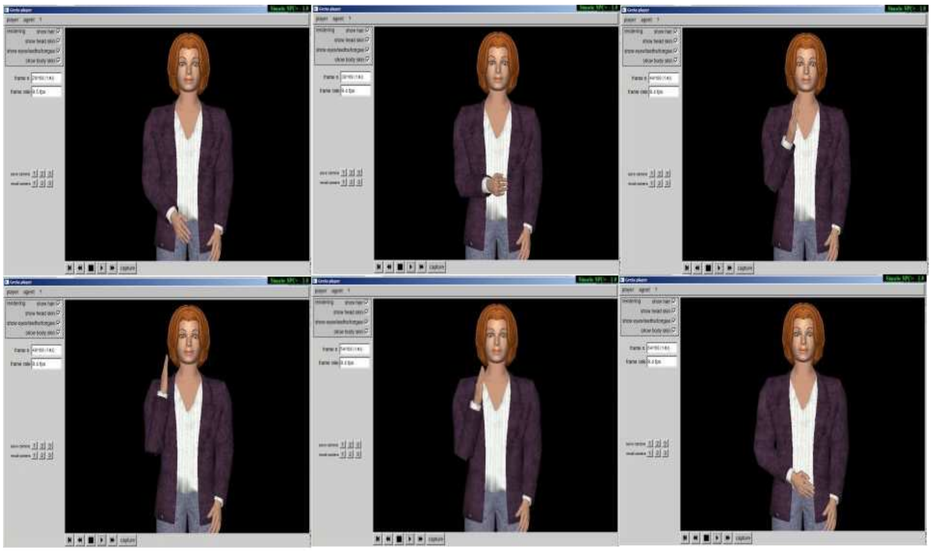


Figure 4-9 Animated gesture using Greta for the spatial extent of V2. Note the small spatial extent.

We synthesize motion trajectories using the Greta animation engine using the temporal extent estimated from videos V1 and V2. The synthesized videos are available on-line[4].

[4] V1TMP: http://www-public.it-sudparis.eu/~rajagopa/BergerG_1234_TMP_03.avi.
V2TMP: http://www-public.it-sudparis.eu/~rajagopa/SimaticG_1234_TMP_02.avi

Figure 4-10 shows the vertical coordinate of the synthesized wrist trajectories. With a higher temporal extent, the trajectory attains its peak and finishes the stroke earlier, and ends earlier. We retrieve the previous result that user V2 has higher temporal extent than user in V1.

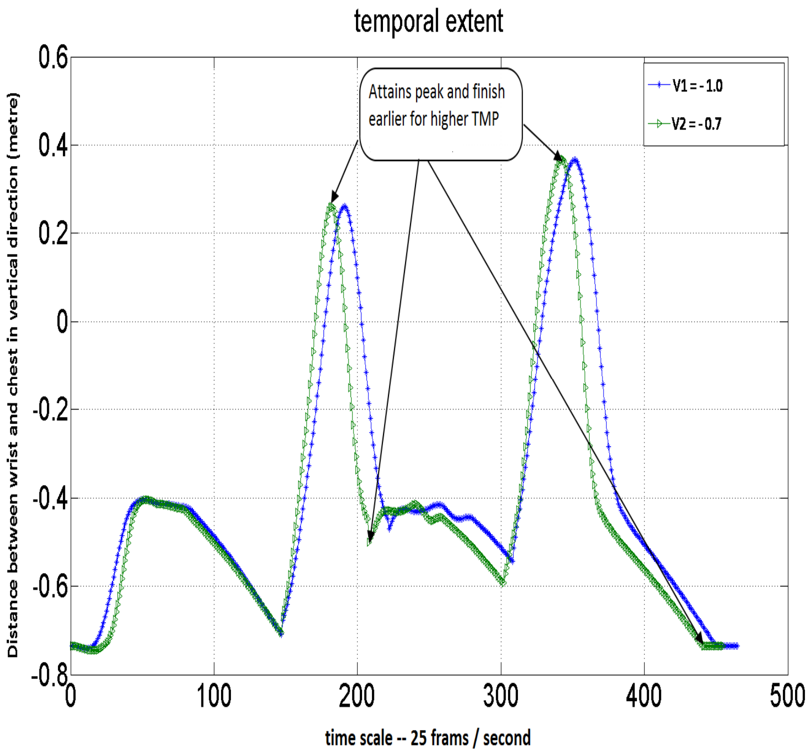


Figure 4-10 Wrist trajectories synthesized using the Greta animation engine, with the temporal extent from V1 and V2.

Similarly we animate the Greta with estimated repetition parameter from videos V2,V3 and V4. Wrist motion trajectories for repetitive actions in the generated animations are shown in Figure 4-11. The generated videos also uploaded in the webpage[5].

[5] http://www-public.it-sudparis.eu/~rajagopa/repVideosForSurvey.htm

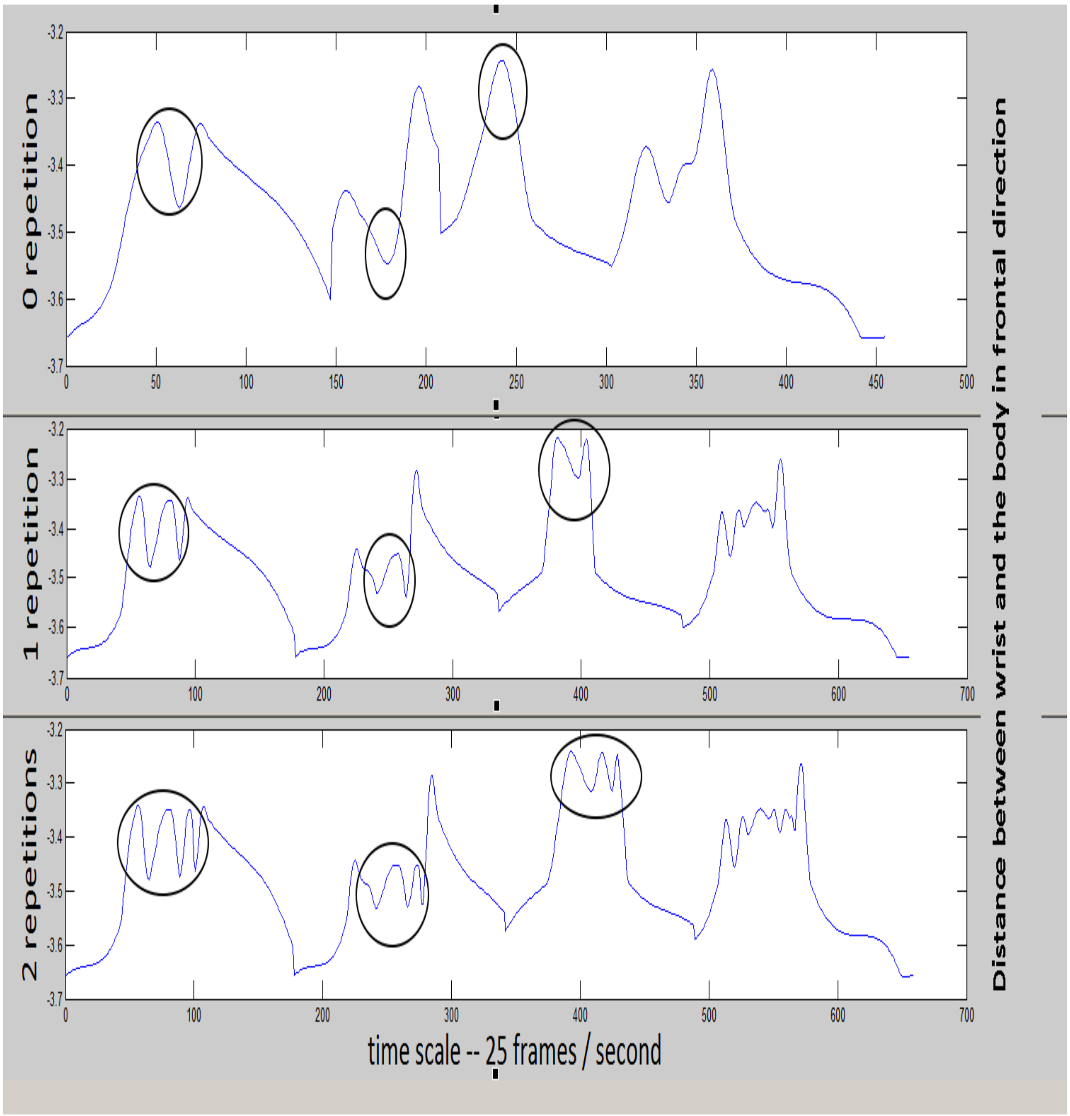


Figure 4-11 : Stroke repetitions in wrist trajectory. Here, only the frontal coordinate of the wrist 3D trajectory is plotted.

4.2 User Reviews

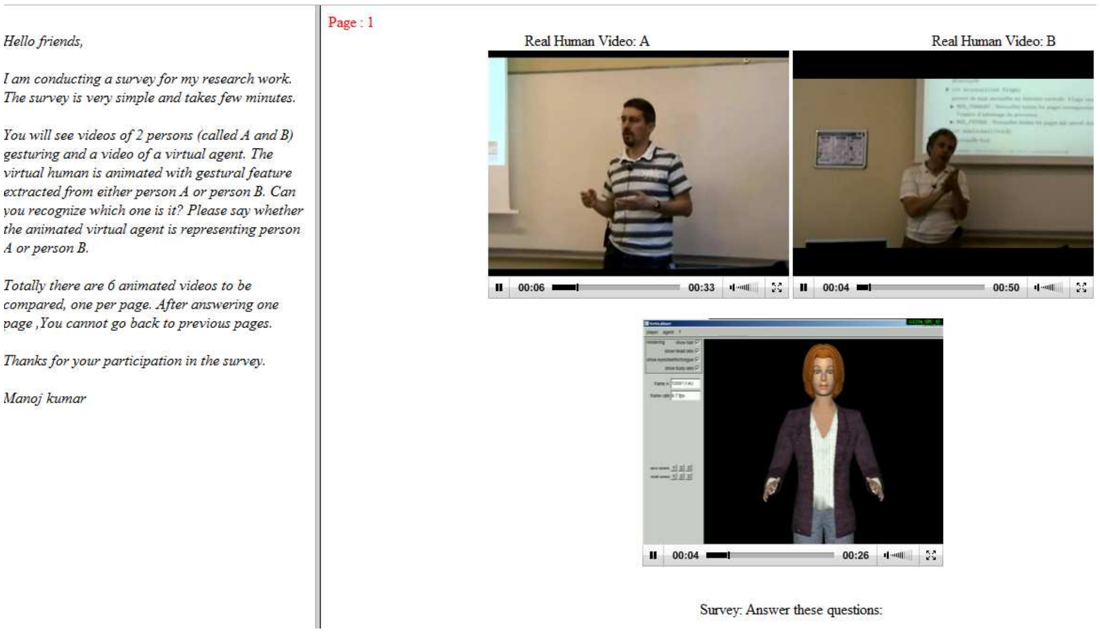
We aim at evaluating the importance of expressivity for virtually cloning human beings. As described previously, we have animated the Greta virtual agent using the expressivity estimated from individual humans. We have conducted an online survey with users that were asked whether they can recognize the individual human behind various synthesis animations.

4.2.1 Review Setup

Users are first asked six personal questions, as follows.

i) In order to check answers uniqueness while protecting user anonymity, we asked users to enter their name through initials as first letter of user's first name, first letter of first name of user's father and first letter of first letter of user's mother.
ii) Gender
iii) Age
iv) Country where the user lived longest
v) User's highest level of education
vi) User's professional field / study

We have conducted reviews in two setups. For both setups, the user receives the following instruction: *"You will see videos of 2 (or 3) persons (called A, B and C) gesturing and a video of a virtual agent. The virtual human is animated with gestural feature extracted from either person A, B or C. Can you recognize which one is it? Please say whether the animated virtual agent is representing person A or person B or people C. Totally there are 6 (or 3) animated videos to be compared, one per page. After answering one page, you cannot go back to previous pages."* For both setups, users are asked: "*which movie is the animated video similar to?*" the answering options are A, B (and C).



Hello friends,

I am conducting a survey for my research work. The survey is very simple and takes few minutes.

You will see videos of 2 persons (called A and B) gesturing and a video of a virtual agent. The virtual human is animated with gestural feature extracted from either person A or person B. Can you recognize which one is it? Please say whether the animated virtual agent is representing person A or person B.

Totally there are 6 animated videos to be compared, one per page. After answering one page ,You cannot go back to previous pages.

Thanks for your participation in the survey.

Manoj kumar

Page : 1

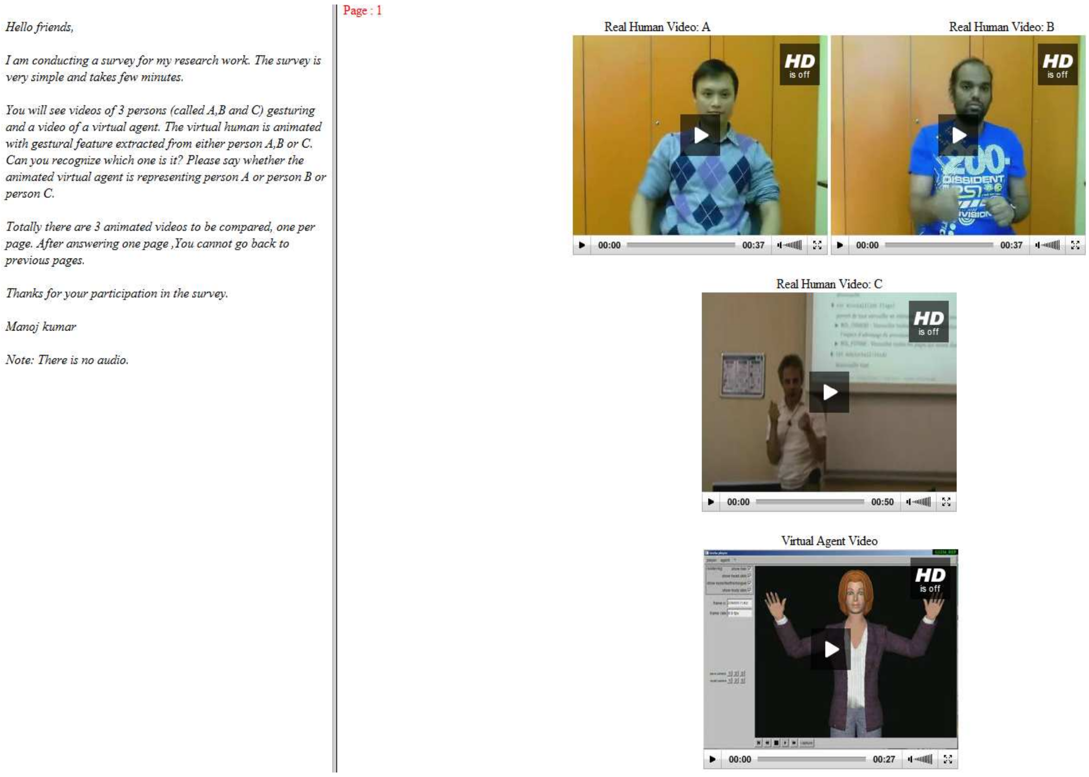
Real Human Video: A

Real Human Video: B

Survey: Answer these questions:

Figure 4-12: Example page from the first user survey setup

Screen shot of the first review setup is shown in Figure 4-12. In the first setup, the V1 and V2 videos are shown against a series of 6 synthesis videos of the Greta virtual agent. In the first and second synthesis videos, the Greta is gesturing with the spatial extent estimated from either V1 or V2. In the third and fourth videos, Greta is animated with temporal extent from either V1 or V2. In the fifth and sixth videos, Greta is animated with both the spatial extent and temporal extent estimated from either V1 or V2.



Hello friends,

I am conducting a survey for my research work. The survey is very simple and takes few minutes.

You will see videos of 3 persons (called A,B and C) gesturing and a video of a virtual agent. The virtual human is animated with gestural feature extracted from either person A,B or C. Can you recognize which one is it? Please say whether the animated virtual agent is representing person A or person B or person C.

Totally there are 3 animated videos to be compared, one per page. After answering one page ,You cannot go back to previous pages.

Thanks for your participation in the survey.

Manoj kumar

Note: There is no audio.

Figure 4-13 Example page from the second user survey setup

Screen shot of the first review setup is shown in Figure 4-13. In the second review setup, the three V2, V3 and V4 videos are presented all together against a series of 3 Greta videos. V2 has no repetitive actions, while gestures in V3 (respectively V4) are repeated once (respectively twice). Input for the Greta animation is 0, 1 or 2 repetitions (in mixed order).

4.2.2 Analysis of User Responses

The statistics of user responses are shown in Table 4-1:

	1st review setup	2nd review setup		
	6 questions	1st question	2nd question	3rd question
No of Participants	17	37	31	26
Male	12	29	24	21
Female	5	8	7	5
Countries				
France	5	24	19	15
Morocco	0	7	6	5
Tunisia	2	2	2	2
India	5	1	1	1
Lebanon	1	1	1	1
Romania	2	1	1	1
Turkey	0	1	1	1
Mexico	2	0	0	0
Level of Education				
Completed high School	0	8	7	5
College Discontinued	0	1	1	1
Doing Bachelors degree	0	11	10	9
Doing / completed Masters degree	12	14	10	8
Doing Phd	5	3	3	3
Field of work				
Computing Robotics	7	13	15	14

	or Cognitive Science				
	Mathematics, Physics or Chemistry	4	10	7	5
	Finance or Economics	1	2	1	1
	Medicine or Biology	1	0	0	0
	Other fields	4	12	8	6
Average Age		28.3	24.3	20.7	22.5

Table 4-1 Reviewers statistics

Results for the first setup are shown in Table 4-2. It appears that more than 71 % and less than 88 % people can recognize the expressivity of the real human from virtual human animation.

	1st test	2nd test	3rd test	4th test	5th test	6th test
Number of participants	17	17	17	17	17	17
Number of users who correctly recognized the real human from the Greta animation	71%	82%	76 %	82%	82%	88 %

Table 4-2 First review setup results

In the second setup, an animated Greta video is compared with 3 real human videos (labeled A, B and C). Results are shown in Table 4-3:

Repetition input to the animation engine	0	1	2
Number of users who correctly recognized the real human from the Greta animation	6 %	41 %	54 %

Table 4-3 Second review setup results

These results show that a high number of repetitions are relevant feature for users to match the animation with the human. Instead, an animation with no repetition would be matched with any human. So the number of repetitions contributes for the recognition.

4.3 Conclusion

In this chapter, we have described our user survey and its results on recognizing real human motion from synthesis motion.

From our first review setup, we found that users could recognize the real human based on the spatial and temporal extents in 75 to 88 % of the trials. This shows the importance of spatial and temporal extent.

The second review setup results show that repetition is not so discriminative. In case gestures are repeated once or twice in the animation, this is perceived as a distinctive feature to recognize humans doing so. Animations with no repetition would be matched with any human, regardless whether they repeat gesture or not.

Chapter 5 - Conclusion

In this work, we address the problem of estimating the style and emotion of a real human. Our approach estimates three parameters that characterize the expressivity of a person. Expressivity parameters were implemented for ECA animation (Hartmann, Mancini, & Pelachaud, 2005), we extended this approach and estimate three parameters from a real human to define the expressivity of a person. Using those expressivity parameters we animate an avatar. The performance of our approach is experimentally validated against synthesis motion and practically tested through user tests. The proposed methods allow estimating some expressivity parameters of a person. In the next sections we summarize the contributions of the thesis and present some future perspectives.

5.1 Summary of Contributions

We have proposed a method to estimate the gesturing expressivity of a person from wrist 3D trajectories. Our contribution consists of: 1) estimating the gesturing expressivity from wrist 3D trajectories and 2) a perceptual study of the relevance of expressivity for recognizing persons.

We estimate three expressivity parameters, namely spatial extent, temporal extent and repetition, from wrist 3D trajectories. For estimating the spatial and temporal extents, we generated the four different hand motion corpora. Out of four different corpora two of them are reference corpora and the remaining two are test corpora. These corpora contain communicative gestures. In the first reference corpora, spatial and temporal extents are varied as explained in the section 3.4.1.1. In the second reference corpus repetition is varied as 0, 1 and 2. Our estimation processes for three parameters are explained in sections 3.4.1, 3.4.2 and 3.4.3.We validated our estimation process with test corpora.

We have animated a virtual agent using the expressivity estimated from individual humans, and users have been asked whether they can recognize the individual humans behind animations. Using estimated expressivity from real human we animated the Greta ECA. We have conducted an online survey with users that were asked whether they can recognize the individual human behind various synthesis animations. We found that, in case gestures are repeated in the animation, this is perceived by users as a discriminative feature to recognize humans, while the absence of repetition would be matched with any human, regardless whether they repeat gesture or not. More importantly, we found that 75 % or more of users could recognize the real human (out of two proposed) from an animated virtual avatar based only on the spatial and temporal extents.

We achieve the virtual cloning by bringing the expressivity of a real human in the virtual human.

5.2 Perspectives

Future research will focus on estimating other expressivity parameters and feed those parameters to the animation engine. For example, the fluidity parameter controls continuity in the transitions between the gesture phases. The power parameter controls hand shapes: high power will shrink the hand (Hartmann, Mancini, & Pelachaud, 2005).

In the user survey, we used 3D motion data captured from monocular images. As mentioned in chapter 4, a real-time 3D sensor, either time-of-flight (Mesa Imaging AG, 2008) (SoftKinetic, 2011) or active triangulation (Microsoft, 2010), can be used for 3D motion capture. Such a sensor achieves more robust motion capture than monocular vision, especially in case of fast motion, potentially leading to a better estimation of expressivity. (Microsoft, 2010)

Finally, a future potential application includes animation movies. A virtual actor can be animated with the expressivity of a real human actor. As an *e.g.* James Stewart and Cary Grant are no more alive, but when we have the expressivity of those actors we can reflect their acting skills in the animation movies.

Another application is virtual embassy. An embassy cannot be present in all the places. The idea is to open the office in different places to answer people queries without real human representative. If an animated character with expressivity of an embassy person is present and answer the queries, then the user will have an experience of talking with a real embassy representative.

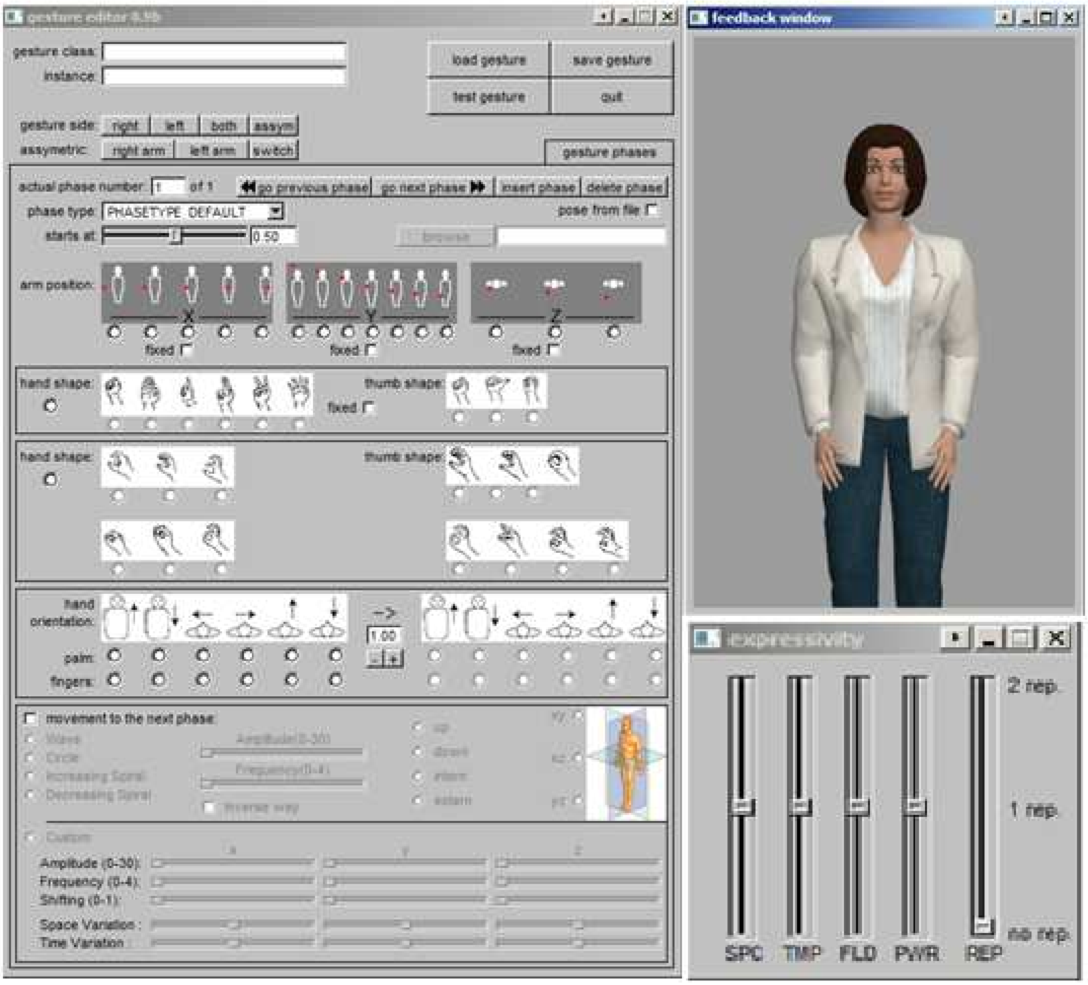
Appendix A

Greta Editors and Players

We used the gesture editors and player that are available with the Greta embodied conversational agent (Hartmann, Mancini, & Pelachaud, 2002).

The editor allows to generate a 3D hand motion trajectories with expressivity and to visualize the motion. The BML (Behavior Mark-up Language) editor was used to merge gestures into a motion. The Greta Player displays the 3D motion.

Here, snapshots of their interface are shown below.



actual phase number: 1 of 1

phase type: PHASETYPE_DEFAULT

starts at: 0.50

Figure A-1: Gesture Editor with expressivity window and visualization window (Hartmann, Mancini, & Pelachaud, 2002)

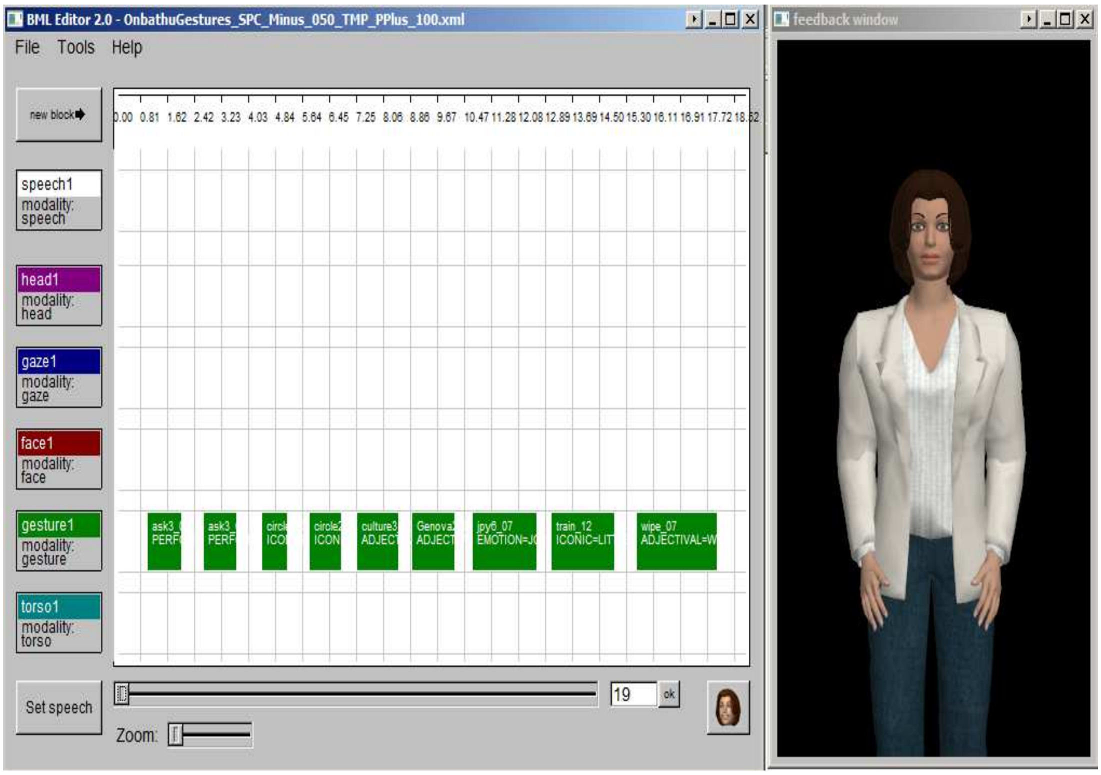
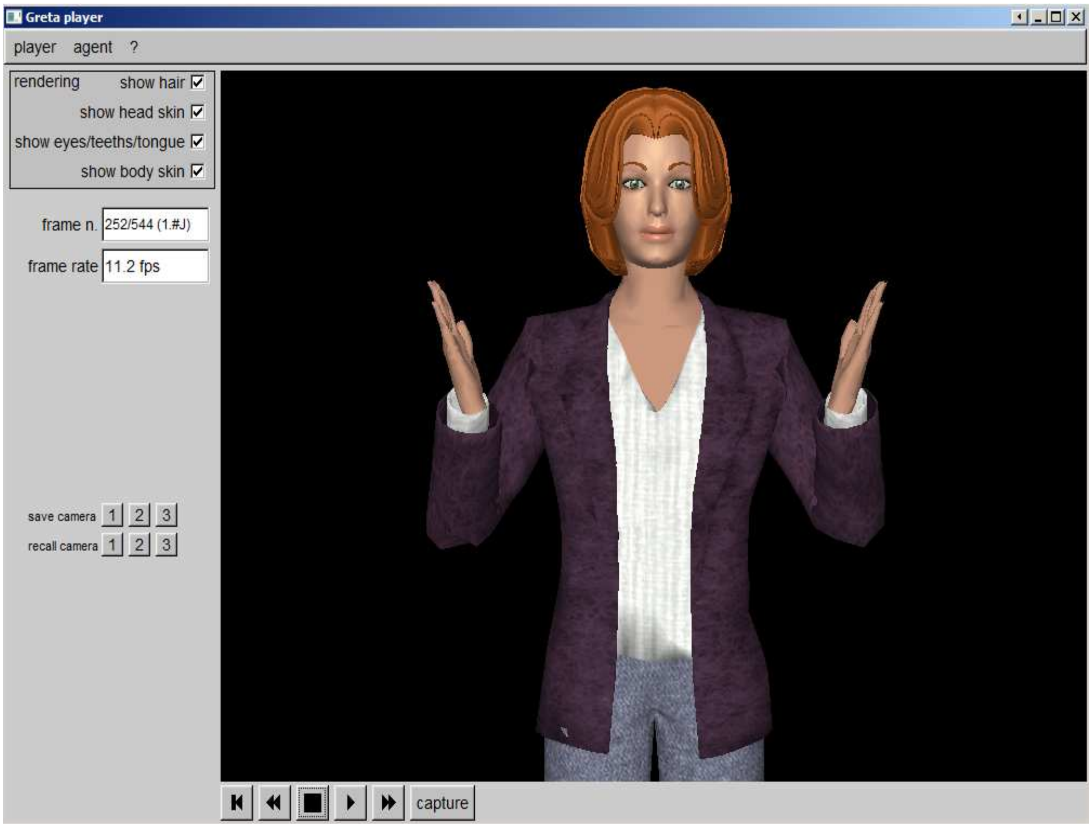

BML Editor 2.0 - OnbathuGestures_SPC_Minus_050_TMP_PPlus_100.xml

Figure A-2: BML Editor (Hartmann, Mancini, & Pelachaud, 2002)


Greta player

rendering show hair ☑

show head skin ☑

show eyes/teeths/tongue ☑

show body skin ☑

frame n. 252/544 (1.#J)

frame rate 11.2 fps

Figure A-3: Greta Player (Hartmann, Mancini, & Pelachaud, 2002)

Publications:

- **Manoj Kumar Rajagopal**, Patrick Horain, Catherine Pelachaud, "Animating a conversational agent with user expressivity", Proc. 11th Int. Conf. on Intelligent Virtual Agents (IVA 2011), Reykjavik, Iceland, September 15-17, 2011, Vol.6895, Pages 464-465.

- **Manoj Kumar Rajagopal**, Patrick Horain, Catherine Pelachaud, "Virtually Cloning Real Human with Motion Style", Proceedings 3rd Int. Conf. on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August 29-31, 2011.

- David Antonio Gomez Jauregui, Patrick Horain, **Manoj Kumar Rajagopal**, Senanayak Sesh Kumar Karri. *"Real-Time Particle Filtering with Heuristics for 3D Motion Capture by Monocular Vision"*, IEEE International Workshop on Multimedia Signal Processing 2010 (MMSP'10), Saint-Malo, France, October 4-6, 2010.

- Donald Glowinski, Matei Mancas, P. Brunet, F. Cavallero, C. Machy, P. J. Maes, S. Passchalidou, **Manoj Kumar Rajagopal**, S. Schibeci, L. Vincze, and G. Volpe, "Toward a model of computational attention based on expressive behavior: applications to cultural heritage scenarios", A. Camurri, M. Mancini, G. Volpe (Eds.), 2010. Proc. 5th Int. Summer Workshop on Multimodal Interfaces (eNTERFACE'09), DIST-Univ. of Genova, Genova, Italy. pp 71-78.

- Matei Mancas, Donald Glowinski, P. Brunet, F. Caveller, C. Mach, P-J. Maes, S. Paschalidou, **Manoj Kumar Rajagopal**, S. Schibeci, L. Vincze, G. Volpe, "Hypersocial museum: addressing the social interaction challenge with museumscenarios and attention-based approaches", QPSR of the numediart research program, Vol. 2, No. 3 , September 2009, pages 91-96.

References

IMVU, Inc. (n.d.). Retrieved from http://www.imvu.com/

Agarwal, A., & Triggs, B. (2006). *Recovering 3D human pose from monocular images.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 28, pp. 44-58.

Allport, G. W., & Vernon, P. E. (1933). *Studies in expressive movements.* The Macmillan Company.

AMM. (2010). *AMM 3D Golf Electromagnetic System*. Retrieved from Advanced Motion Measurement, Inc.: http://www.advancedmotionmeasurement.com/Products/AMM3DElectromagneticSolution.aspx

Arikan, O., & Forsyth, D. (2002). Interactive motion generation from examples. *SIGGRAPH 2002, 21*, pp. 483-490.

Badler, N. I., Palmer, M. S., & Bindiganavale, R. (1999). Animation control for real-time virtual humans. *Communications of the ACM , 42* (8).

Ball, G., & Breese, J. (2000). *Embodied conversational agents.* MITpress.

Bänziger, T., Pirker, H., & Scherer, K. R. (2006). GEMEP – GEneva Multimodal Emotion Portrayals:A corpus for the study of multimodal emotional expressions. *5th International Conference on Language Resources and Evaluation (LREC 2006).* Genova, Italy.

Bevacqua, E., Mancini, M., Niewiadomski, R., & Pelachaud, C. (2007). An expressive ECA showing complex emotions. *Proceedings of the AISB Annual Convention*, (pp. 208-216). Newcastle,UK.

Black, A. W., Taylor, P., Caley, R., & Clark, R. (n.d.). *Festival*. Retrieved from http://www.cstr.ed.ac.uk/projects/festival/

Boone, R. T., & Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: the development of cue attunement. *Developmental Psychology , 34* (5), 1007-1016.

Brand, M., & Hertzmann, A. (2000). Style Machines. *SIGGRAPH 2000*, (pp. 183-192).

Calvert, T. W., & Chapman, A. E. (1994). Analysis and synthesis of human movement. In *Handbook of Pattern Recognition and Image Processing* (Vol. 2, pp. 432-474). San Diego , USA.

Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., et al. (2004). Toward real-time multimodal processing: EyesWeb 4.0. *Proceedings of AISB 2004 Convention:Motion, Emotion and Cognition.* Leeds , UK.

Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies, Elsevier Science , 59* (1-2), 213-225.

Caridakis, G., Raouzaiou, A., Bevacqua, E., Mancini, M., Karpouzis, K., Malatesta, L., et al. (2008). Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation , Special issue on Multimodal Corpora For Modelling Human Multimodal Behavior* , 367-388.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. F. (2000). *Embodied Conversational Agents.* MIT Press.

Castellano, G., & Mancini, M. (2009). Analysis of Emotional Gestures for the Generation of Expressive Copying Behaviour in an Embodied Agent. In *GESTURE-BASED HUMAN-COMPUTER INTERACTION AND SIMULATION* (pp. 193-198). Berlin / Heidelberg: Springer.

Chellappa, R., Roy-Chowdhury, A. K., & Kale, A. (2007). Human Identification using Gait and Face. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (pp. 1-2). Minneapolis, MN, USA.

Chi, D., Costa, M., Zhao, L., & Badler, N. (2000). The EMOTE Model for Effort and Shape. *SIGGRAPH 2000,* (pp. 173-182).

Craig, J. J. (1986). Forward Kinematics. In *Introduction to robotics : mechanics and control* (3rd Edition ed.). Prentice hall.

Efron, D. (1941). *Gesture and environment.* New York: King's Crown Press.

Ekinci, M. (2006). A New Approach for Human Identification Using Gait Recognition. *ICCSA, 3,* pp. 1216-1225. Glasgow, UK.

Ekman, P. (1982). *Emotion in the human face.* Cambridge University Press.

Elgammal, A., & Lee, C. S. (2004). Separating style and content on a nonlinear manifold. *Computer Vision and Pattern Recognition (CVPR),* (pp. 478-485).

Eliëns, A., Huang, Z., & Visser, C. (2002). A platform for Embodied Conversational Agents based on Distributed Logic Programming. *AAMAS Workshop on "Embodied conversational agents – Let's specify and evaluate them!".* Bologna,Italy.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude. *Journal of Experimental Psychology , 47* (6), 381-391.

Furniss, M. (2000). *MOTION CAPTURE: AN OVERVIEW.* Retrieved from http://www.animationjournal.com: http://www.animationjournal.com/abstracts/essays/mocap.html

Gallaher, P. E. (1992). Individual Differences in Nonverbal Behavior: Dimensions of Style. *Journal of Personality and Social Psychology , 63* (1), 133-145.

Gallistel, C. R. (1980). *The Organization of Action : A New Synthesis.* Lawrence Erlbaum Associates.

Gómez Jáuregui, D. A., Horain, P., Rajagopal, M. k., & Karri, S. S. (2010). Real-Time Particle Filtering with Heuristics for 3D Motion Capture by Monocular Vision. *Multimedia Signal Processing*, (pp. 139-144). Saint-Malo, France.

Gonzalez, R. C., & Woods, R. E. (1992). *Digital Image Processing.* Pearson Prentice Hall.

Grochow, K., Martin, S. L., Hertzmann, A., & Popović, Z. (2004). Style-based Inverse Kinematics. *ACM Transactions on Graphics(Proceedings of ACM SIGGRAPH 2004) , 23* (3), 522-531.

Gypsy7. (2010). *Animazoo motion capture system and technology*. Retrieved from http://www.animazoo.com/motion-capture-systems/gypsy-7-motion-capture-system/

Hartmann, B., Mancini, M., & Pelachaud, C. (2002). Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis. *Computer Animation.* Geneva: IEEE Computer Society Press.

Hartmann, B., Mancini, M., & Pelachaud, C. (2005). Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. In *Gesture Workshop,LNAI.* Springer.

Hassin, R. R., James, U. S., & Bargh, J. A. (Eds.). (2005). *The New Unconcious.* Oxford University Press.

Hsu, E., Pulli, K., & Popović, J. (2005). Style translation for human motion. *ACM Transactions on Graphics , 24* (3), 1082-1089.

IGS-180. (2010). *Inertial gyroscopic motion capture system*. Retrieved from Animazoo motion capture system and technology: http://www.animazoo.com/motion-capture-systems/igs-180-motion-capture-system/

Intersense. (2010). *Intersense IS-900, Sensing every move.* Retrieved from Intersense Inc. Billerica, MA, USA: http://www.intersense.com/IS-900_Systems.aspx

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *14* (2), 201-211.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature , 290*, 91-97.

Kendon, A. (2004). *Gesture Visible action as utterance.* Cambridge University Press.

Kinect, M. (2010). *Microsoft Kinect*. Retrieved from http://www.xbox.com//kinect

Kita, S., Gijn, I. v., & Hulst, H. v. (1998). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. *International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, (pp. 23-35).

Klima, E., & Bellugi, U. (1979). *The signs of language.* Harvard University Press.

Kopp, S., Sowa, T., & Wachsmuth, I. (2003). Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. *Gesture Wrokshop*, (pp. 436-447).

Kovar, L., Gleicher, M., & Pighin, F. (2002). Motion Graphs. *SIGGRAPH 2002*, (pp. 473-482). 21.

Laban, R. V. (1960). *The Mastery of Movement.* London: MacDonald & Evans.

Lee, J., Chai, J., Reitsma, P. S., Hodgins, J. K., & Pollard, N. S. (2002). Interactive control of avatars animated with human motion data. *SIGGRAPH 2002*, (pp. 491-500).

Lee, L., & Grimson, E. (2002). Gait Analysis for Recognition and Classification. *Fifth IEEE International Conference on Automatic Face Gesture Recognition*, (pp. 734-742). Washington,USA.

Li, Y., Wang, T., & Shum, H.-Y. (2002). Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis. *SIGGRAPH 2002*, (pp. 465-472).

Linden Research, Inc. (n.d.). Retrieved from http://secondlife.com/

Mancini, M., Bresin, R., & Pelachaud, C. (2007). A virtual-agent head driven by musical performance. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, *15*, pp. 1883-1841.

McNeill, D. (1992). *Hand and Mind what gestures reveal about thought.* Chicago, USA: The university press of chicago press.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *6*, 109-114.

Meijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior , 13* (4), 247-268.

Mesa Imaging AG. (2008). *SwissRanger SR4000 - miniature 3D time-of-flight range camera.* Retrieved January 21, 2012, from Mesa: http://www.mesa-imaging.ch/TechOverView.php

Microsoft corporation. (n.d.). *Kinect*. Retrieved November 2010, from http://www.xbox.com/en-GB/kinect

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *International Journal Computer Vision and Image Understanding (CVIU'06) , 104* (2), 90-126.

Niewiadomski, R., Hyniewska, S., & Pelachaud, C. (2009). Modeling emotional expressions as sequences of behaviors. *International conference on Intelligent virtual agents (IVA)*, *5773*, pp. 316-322. Amsterdam.

Noot, H., & Ruttkay, Z. (2005). The GESTYLE Language. *International Journal of Human-Computer Studies - Special issue: Subtle expressivity for characters and robots , 62* (2).

Optitrack. (2010). *Optitrack-optic track motion capture system*. Retrieved from http://www.naturalpoint.com/optitrack/

Pelachaud, C. (2009). Modelling multimodal expression of emotion in a Virtual Agent. *Philosophical Transactions of Royal Society B Biological Science* (364), 3539-3548.

Pelachaud, C. (2005). Multimodal expressive embodied conversational agent. *ACM Multimedia*, (pp. 683-689). Singapore.

Pelachaud, C., & Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation , 13* (5), 301-312.

Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication.* Weidler Buchverlag Berlin.

Poggi, I. (2001). Towards the lexicon and Alphabet of Gesture, Gaze, and Touch. *Virtual Symposium http://www.semioticon.com.*

Poggi, I., & Pelachaud, C. (2000). Emotional Meaning and Expression in Animated Faces. In A. Paiva (Ed.), *Affective interactions: towards a new generation of computer interfaces.* Berlin: Springer-Verilag.

Poggi, I., & Pelachaud, C. (2008). *Persuasion and the expressivity of gestures in humans and machines.* (I. Wachsmuth, M. Lenzen, & G. Knoblich, Eds.) Oxford University Press.

Polana, R., & Nelson, R. (1994). Low level recognition of human motion. *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, (pp. 77-82). Ausitin.

Pollick, F. E. (2003). The features people use to recognize human movement style. (A. Camurri, & G. Volpe, Eds.) *Gesture-Based Communication in Human-Computer Interaction* , 10-19.

Poppe, R. (2007, October). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding* , 4-18.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., & Henning, J. (1989). *HamNoSys. Version 2.0. Hamburg Notation System for Sign Languages:An Introductory Guide.* Hamburg, Germany: Amburg Signum.

Quek, F. K. (1995). Eyes in the interface. *Image Vision Computing , 13* (6), 511-525.

Quek, F., McNeill, D., Bryll, R., Kirbas, C., & Arslan, H. (2000). Gesture, speech, and gaze cues for discourse segmentation. *Computer Vision and Pattern Recoginition(CVPR), 2*, pp. 247-254.

Reinhard, E., Ashikhmin, M., Gooch, B., & Shirley, P. (2001). Color Transfer between Images. *IEEE Computer Graphics and Applications , 21* (5), 34-41.

SoftKinetic. (2011, December 20). *Announces First Affordable Time-of-Flight Depth-Sensing Camera for Commercial Use.* Retrieved January 21, 2012, from SoftKinetic: http://www.softkinetic.com/PressRoom/News/tabid/322/ArticleId/107/SoftKinetic-Announces-First-Affordable-Time-of-Flight-Depth-Sensing-Camera-for-Commercial-Use.aspx

Stokoe, W. (1960). *Sign Language Structure.* Occasional Paper, Studies in Linguistics.

Tenanbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Computation , 12* (6), 1247-1283.

Thalman, D. (2000). Challenges for the Research in Virtual Humans. *Workshop on Achieving Human-Like Behaviour in Interactive Animated Agents, AGENTS 2000.* Barcelona, Spain.

Thalmann, N.-M., & Thalmann, D. (1991). Complex models for animating synthetic actors. *IEEE Computer Graphics and Applications, 11(5)*, pp. 32-44.

Thalmann, N.-M., Jain, L. C., & Ichalkaranje, N. (Eds.). (2008). *New Advances in Virtual Humans: Artificial Intelligence Environment.* springer.

Thrun, S., & Pratt, L. (1998). *Learning To Learn.* Norewell,MA,USA.: Kluwer Academic Publishers.

Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3d people tracking with gaussian process dynamical models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1*, pp. 238--245.

Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., & Fua, P. (2004). Style-Based Motion Synthesis. *Computer Graphics Forum , 23* (4), 799-812.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology , 28* (6), 879-896.

Wallbott, H. G. (1985). Hand movement quality: A neglected aspect of nonverbal behavior in clinical judgment and person perception. *Journal of Clinical Psychology , 41* (3), 345-359.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology , 51* (4), 690-699.

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2008). Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, pp. 283-298.

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2007). Multifactor Gaussian Process Models for Style-Content Seperation. *ICML*. Corvallis, OR , USA.

Wu, X., Ma, L., Zheng, C., Chen, Y., & Huang, K.-S. (2006). On-line motion style transfer. *International Conference on Entertainment Computing. 4161*, pp. 268-279. Cambridge , UK: LNCS.

Zhao, L. (2001). *Synthesis and Acquisition of Laban ovement Analysis Qualitative Parameters for Communicative Gestures.* University of Pennsylvania.

Zhu, J., & Thagard, P. (2002). Emotion and action. *Philosophical Psychology , 15* (1), 19-36.