Capturing Human Nuance

Timothy Campbell Lukins

PhD. Proposal

Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh

Supervised by:
Prof. R.B. Fisher - IPAB, University of Edinburgh.
Dr. C. Urquhart - Virtual Clones Ltd.
Dr. J. Oberlander - ICCS, University of Edinburgh.

Abstract

Current computer vision and graphics techniques for modelling human dynamics have a wide range of applications for activity recognition, gesture tracking, and animation. While successful in many respects, they have issues with interpreting and generating subtle actions that appear similar, yet have completely different meanings to a human observer. That is, they fail to account for the additional information inherent in *nuance* - the slight variations in dynamic features that serve to define realistic and recognisable expressive modes by altering the basis of a common underlying movement. Some effort has in recent years been based in exploiting motion-capture data for the similar idea of separating style from structure. This prior work is however biased to pre-selected joint and feature based sets, does not integrate well with other research into facial expression recognition and modelling, and lacks application of other applicable machine-learning techniques. In this proposed research, we focus on the specific problem of extrapolating the aspects of variation that define nuance via a more holistic treatment of a dense 3D temporal-spatial data. In particular, by unsupervised analysis of range-flow to provide more accurate and meaningful dynamic features. From this we intend to decouple a statistical model, and to correlate this within an established psychological framework (FACS). This will hopefully advance understanding, classification and synthesis of the subtler aspects of human expression.

Contents

1 Introduction

The increasing sophistication of advanced computational models for representing humans, both in computer vision (for tracking, activity recognition, shape recovery) and in computer graphics (for behavioural animation, motion synthesis, virtual capture), demand great effort and novel approaches in order to generate good results. This problem is on one hand a factor of the complexity in accommodating all the *spatial* aspects for alterations in the forms that the human body is capable of, which even for a fairly simplified articulated model or surface soon becomes increasingly complex. It is also a problem in accommodating the effectual subset of *temporal* sequences that reflect natural motion dynamics that serve to express a wide range of particular meanings dependant on the context and observer. In combination, these together raises the fundamental questions of what combinations of *temporal-spatial* features of the human body serve to define realistic behaviours and actions? Can such features be isolated and used as the basis for defining classes of dynamics? Which of the these features may be responsible for triggering aspects of the human perceptive system used for recognition of emotion and expression?

Currently, for example, if an animator wishes to impart a particular sense of believability and vitality to a virtual actor in a motion film, hours must be spent tweaking and explicitly controlling the raw model to achieve the desired result such that the character becomes “alive”. While many tools and techniques exist for expediting this manipulation, it still requires great skill and characterisation to perform. It is furthermore made all the harder if different components interact at the same time, or with a desired dramatic motivation - to the extent that the process may well be guided by a filmed actor who is traced over and supplanted (a technique exploited in the recent Lord of the Rings trilogy to great effect).

Conversely, the ability to be able to differentiate between ever more slight and similar gestures or motions is a focus of much research in computer vision for diverse applications such as security monitoring, sign language interpretation, and gesture control. This is achieved mostly from building a suitable classification model from training sets constructed with hours of painstaking work in hand annotating video footage, with a certain level of loss as a result in the resolution and 2D nature of such data and further limited viewpoint. Even with fairly accurate 3D information gained from motion-capture based techniques there is a fundamental issue in constructing a reliable statistical model from a feature set that does not necessarily encompass the relevant information, and is biased to the placement of sensors.

In both graphics and vision there is consequently a need for a modelling process that can acquire and accommodate those elements of human dynamics that define realism and distinguish between the subtleties of differing acts. This idea points toward some means of analysing the variation that dictates recognisable forms of human expression occurring across different dimensions. Having access to such an abstraction would thus hopefully allow control of an associated model, such that animation could be enhanced simply by specifying the range of allowable dynamics (e.g. a walk sequence would be specified by single or mixture of modes like “stealthy", “hesitant", etc.) Similarly, vision systems would be able to actually recognise such modes on the basis of a model that is able able to isolate the elements that distinguish them.

In the remainder of this opening section we present a high-level overview of the basis for this research, to effectively *define* the problem more explicitly and the *contribution* we will make by investigating it. The rest of this proposal is then structured into two remaining chapters to describe in detail how we hope to achieve this. Firstly, to *review* the fields that touch on the concept of nuance and what it involves - building on the basis of what is known about the subtlety of human dynamics and how they are understood, represented, and captured. This leads secondly to describing the *approach* we shall adopt to further the research by focusing on the problem - as defined by our key hypothesis below and how we shall be able to construct a framework in order to experimentally test for its validity.

1.1 Definition

The nature of human dynamics is superficially simple when observed from our own viewpoint. This is primarily because our visual and cognitive understanding - facilitated by our evolution and development - enables us to detect, recognise, anticipate, and interact with other people in a way that we take for granted during our day-to-day existence. Because it is such a natural part of our experience, we are all adept as individual human motion and shape “detectors” and rely on our internal representations to constantly focus, analyse and provide us with information about what those others out there in the world are doing or going to do, what they are feeling, what they look like, what their response is to us, etc.

This presents a major problem when attempting to construct realistic artificial models, since because we are so attuned to perceiving real people - we are consequently equally as good in spotting errors with these aspects when then go wrong. Those flaws which for one reason or another do not hold true to our understanding and violate our expectations, such that the closer any representation resembles that of an actual human the more critical we are of it (if we suspect it), and the stronger our aversion is to it if we realise it is fake.

Based on this observation, we seek to emphasise the importance of the concept of ***nuance*** **as the variations in dynamic shape and motion that serve to distinguish purposeful acts of a similar basic form.** A such, a nuanced activity (to quote the OED) is one which *possesses or exhibits delicate graduations in tone, expression, meaning, etc.* - ultimately contributing in defining realistic and recognisable behaviours. Whether such activity is moving the arm to point, generating a different vocalisation, or knotting the brows - we as humans are able to observe key elements, and to associate them with an **intention independent of the individual or the instance**.

From such a definition - and based on the general observations of human movement - we can offer the following initial examples of *aspects of nuance* that, when combined, serve to define different styles, expressions, and even emotions or by their deviation from standard norms:

- *Timing.* The variation in dynamics controlled by velocity and acceleration (not exceeding any limits in being too fast or slow nor too rhythmic).
- *Co-ordination.* The variation in dynamics for possible combinations occurring relatively at the same instance (not appearing too symmetrical or impossible).
- *Extent.* The variation in dynamics within the range of minimum and maximum alteration (not going out with the limits of physical plausibility).
- *Direction.* The variation in dynamics occurring along a particular trajectory, extending from a particular origin position to a final position.

Thus, the sum total effects of nuance acts to regulate dynamics in increasingly complex and variedly expressive form and motion, by contributing in different ways and at different points in a temporal sequence. As such, nuance does not necessarily describe the most efficient way in performing an act - in fact, it most often describes the opposite case in which efficiency is sacrificed in order to convey additional meaning. The act will also vary by a certain amount for every time it is carried out, and also by who is performing it - but crucially retaining those universal features that distinguish it.

The simplest example of nuance in operation can be illustrated by considering the dynamics of a hand. Firstly, as it moves backward and forward in space, its velocity varying accordingly (timing). Secondly, the fingers move independently and concurrently in an arbitrary sequence (co-ordination) and to different degrees (extent). Lastly, the expression for the action can vary to the observer as a “point” or “beckon” depending on the order of movement (direction). In combination these factors can vary to describe a wide range of acts that can be classified such as “aggressive points” and “hesitant points”, or “curt beckons” and “suggestive beckons”.

Similarly, the dynamics of the face represent a very compact yet flexible source of nuance, in that the slightest subtle variation in feature such as the raising of the cheeks and wrinkling of the nose can serve to generate a completely different expression (c.f. “disgust” with “bemused”). However, the face is very much more *punctuated* in terms of its overall dynamics, with most expressions taking only a second or so to form. It is therefore interesting to observe the *transitions* that occur between expressions, especially in the context of a conversation when different features are called upon independent of the articulation of the lips in order to provide subsidiary emphasis at key points.

1.2 Hypothesis

From the above observations and clarification of nuance, we postulate the following hypothesis to express the focus of this research:

> Nuance embodies the variations in features that serve to define realistic and recognisable human expressive modes by altering the basis of common underlying form.

The assumptions raised by this statement are as follows:

- That “*features*” exist as a set of localised regions of the body whose motion and shape provide the basis for various forms of expression.
- That “*variation*s” exist as a range of trajectories of these features along a high-dimensional manifold in the space of all human expressive modes.
- That “*human expressive modes*” exist as styles, emotions, or gestures that can be uniquely recognised as relaying a specific intention or purpose.
- That “*common underlying form*” exist as universal physical dynamics that are independent of individual performance and instance.

1.3 Contribution

The focus of this research builds on the wealth of knowledge that exists in understanding the perception of the human form, how its dynamics can be observed and recorded, and how these can be modelled and interpreted by modern computational means (as detailed in the following review chapter). Within this general research area, we believe that the identification of *nuance* as a key underlying concept for understanding human expression, has not been explicitly proposed before.

In particular, we seek justification for this proposal by taking the following unique approach of:

1. Exploiting **temporal 3D range-data from stereo** as provision for high-resolution input.
2. Extracting features from this rich input via **unsupervised interpretation of range-flow**.
3. Decoupling a **statistical model of the variation that is nuance** from the underlying dynamics.
4. Validating the model by **correlating it within an established psychological framework**.

The proposed research thus aims *to advance understanding by providing a new capture and modelling framework for supporting work in human dynamics, and to use it to validate to what extent nuance plays a role in defining how we perceive realistic expressive behaviou*r. We shall focus on the face (as the most compact source of nuance) - but shall be able also to generalise to the whole body. We hope that this work will in turn provide advances and applications in computer vision, graphics, biometrics and psychophysics.

2 Review

In this chapter we present a *review* of the related research into the perception and modelling of human body dynamics, and the subtleties conveyed by them. The history of this has a long tradition extending back to ancient studies of anatomy and representation of the human form, and are encountered today as a number of diverse modern fields. We present this background as a series of perspectives which aim to focus on those aspects of the problem concerned explicitly with *nuance*. That is, we do not present the intricacies of physiology and neurology, or the low-level detail of various computational models or notational schemes for defining and controlling animation (many good reference books are available for this end [59, 48, 81]).

Our intention instead is to show that we have identified the body of research that represents the core foundations on which this work is based. To describe and highlight the approaches that have already encountered the problem of human nuance, and which have attempted to describe and analyse its properties and effects. No single field explicitly tackles these issues directly, indeed none use the term “nuance” explicitly, which requires us to present the problem as it is encountered in the following key directions:

- The *perception* of human dynamics from the cues provided by nuance - as investigated by research into people’s ability to determine those inherent aspects of activities that contribute to the recognition and realism of the most subtle forms of actions.

- The *description* of human dynamics to the levels of detail required to capture nuance - including the various approaches that have sought to transcribe and automatically capture the finest degree to which subtlety of action can be represented.

- The *interpretation* of human dynamics in order to extrapolate nuance - the focus of work into decoupling motion into those components that represent underlying form and those that constitute the stylistic variations which define subtly different actions.

At the end of this section we form a *critique of nuance*, to act as a summary of where the issues continue to lie and how research might be extended and tackled, as presented in the next chapter.

2.1 Perceiving Nuance

There exists a great deal of understanding of the underlying biology that serves to generate and control natural dynamics. However, another studied area is the inverse problem of how humans then *perceive* and thus differentiate such a vast range of movement. This is a important point: that while we are all capable of generation to some degree or other, it is only through complicity with the observation that an act is distinguished. This idea give rise to the area of study under *perceptual psychophysics*, in seeking to quantify the nature and extent of the human ability to recognise such factors as gender, age, emotion, identity, task, intention, etc. - in response to factors such as the relative scarcity, noise, and exaggeration of the effective signal and image features. Such studies have focused on the very subtle distinctions that facilitate such an ability, for example by relying on “point-light" data, either captured from real people or simulated, in order to quantify to what extent various features play a part in triggering a response (even when presented statically as in Figure 1). The importance in the selection of such features and their temporal-spatial characteristics is an important result of this work in understanding human perception of others.

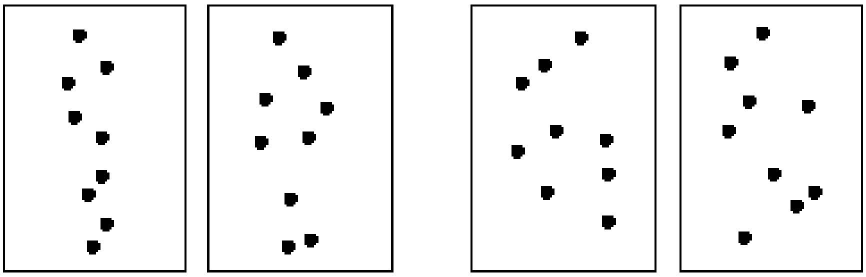


Figure 1: Point light studies for normal and scrambled motions (*Cutting 1977*).

A commonly used concept and experimental foundation used within such studies is that of an ideal observer (originally a term used in the context of signal detectability in engineering) to statistically validate a model of sensitivity to isolated features in the presence of noise. This is done in order to quantify the performance of actual people to different visual tasks when presented with images, by contrasting their responses to those of the ideal. Using such an approach, the ability of humans to robustly and accurately recognise impoverished (e.g. point-light) stimuli is well documented, and forms an active research area. The actual term “biological motion” is derived from those early psychophysical studies performed by Cutting, Kozlowski and Johansson which served to initially identify the human ability to recognise action solely on the basis of motion. A more complete overview of the history of analysing human motion from this perspective, and subsequent use of a derived model for synthesis in testing, is described by Troje [74]. Such experimental work focuses on quantifying exactly what variation and subtlety is permitted in recognition, and the effects of modulating even the most sparse set of features - for example, exaggerating spacial-temporal factors in tennis serves [64].

An analysis of the actual neural mechanisms that surround the ability for recognition and learning to identify motion is presented in Giese and Poggio [37]. Current understanding has lead to the general acceptance of a model for the interpretation of biological motion via two parallel pathways that handle *form* and *motion* respectively - each of which comprise a hierarchy of neural feature detectors that operate at increasing levels of complexity and scale. Thus, at the lowest level small receptive fields are in operation in detecting localised orientation and motion, which are effectively combined to form invariant representations and sequences of snapshots/prototypes of characteristic dynamics (reinforced by lateral connections between the pathways). Recent advances attempt to break away from modelling such low-level descriptions of the human visual system (and the inevitable complexity), and to instead focus on modelling the higher-level cognitive processes within a statistical framework. An example of this is to exploit the process of Bayesian inference to account for the observation (originally made by Helmholtz) that *retinal images are ambiguous, and that prior knowledge is needed to account for perception*. This fits a probabilistic model, with the need for a combination of likelihood of a scene having such image features, and the prior probability of such a scene occurring in the first place as learnt from experience. As described in Kersten *et al.* [44], such a model provides a framework for psychophysical experiments in object recognition at many different levels - in which Bayesian observers provide good approximations of actual human performance.

Attempts to quantify the degree to which the perception of facial expressions can be differentiated have raised some particularly interesting results. The recent work of Young *et al.* construct experiments utilising techniques in computer graphics to "Megamix" faces (Figure 2) in order to assess how much percentage of an expression (70%+) is required to consistently recognise it [85]. Further work also revealed the role of intensity to caricature and thereby enhancing identification, and how configural coding - the relationship between facial features - is also important to the recognition of expression, and is separately processed from the use of features for identity (seen by PCA analysis) [18].

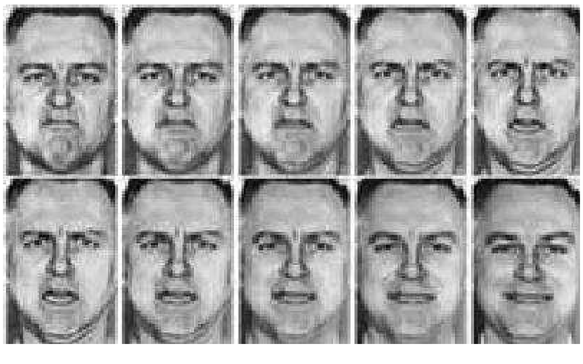


Figure 2: "Megamixing" to determine expressoin identification boundary (*Young et al. 1997*).

Understanding *why* humans move as they do is determined by those evolutionary, sociological, emotional, motivational and behavioural factors that constitute elements of psychology. Studies of these factors are centred on the observation that dynamics are generally shaped by the state of the person and the context of the situation - focusing on describing the actions of a person in such a way as to quantify the underlying purpose (i.e. to derive the state or context). In psychology, this *non verbal communication* has been set out in the seminal work of Argyle [3] as playing a central role in human social behaviour. It is more formally known as *kinesics* - the study of those body movements and gestures by which (as well as by speech) communication is made. Research in this is in turn part of the wider field of *semiotics* - communication studied through the interpretation of symbols). A more specialised overview of this field is presented by Birdwhistell [9], and his work based on the conviction that "*body motion is a learned form of communication, which is patterned within a culture, and which can be broken down into an ordered system of isolable elements*". His approach is reliant on cinematographic techniques to assist in the scientific research into human dynamics, effectively providing a valuable tool with which to be able to attempt to understand and interpret complex nonverbal communications based on touch, posture and gesture. This focus then lies on using a variety of schemes to finely annotate the particular motions of the observed people, with the aim of then being able to analyse and derive an underlying *deeper structure* (a "grammar" - the existence of which is a continuing source of contention with the psychological community), in the hope that having such a scheme will facilitate interpretation of non-volational movement.

Related to the work in kinesics is the large body of investigation conducted by Paul Ekman over the last 40 years specifically into *facial expression*. This work is in turn inspired by the monumental keystone of Charles Darwin's treatise "*The Expression of the Emotions in Man and Animals*" [26] (often neglected because of his work on evolution). It was Darwin who first wrote how 'it has often struck me as a curious fact that so many shades of expression are instantly recognised without any conscious process of analysis on our parts', and to try and evaluate its universality. On this basis, later research has tried to establish the true facts as to the cultural [30] and communicative [29] ramifications. A key by-product of this work is a realisation of the necessity in understanding the formation and variation in different elements that come together to create a final expression - and to be able to describe how they do so. Initially, Ekman proposed that there were effectively six basic forms existing at the extreme of various positions, corresponding to the basic extremes of emotions (anger, fear, surprise, disgust, joy, sadness - Figure 3). However, it was soon realised that this formed an approach that was unable to handle the diversity and ambiguity inherent to the variation in face shape and individual performances - instead requiring a more detailed approach.

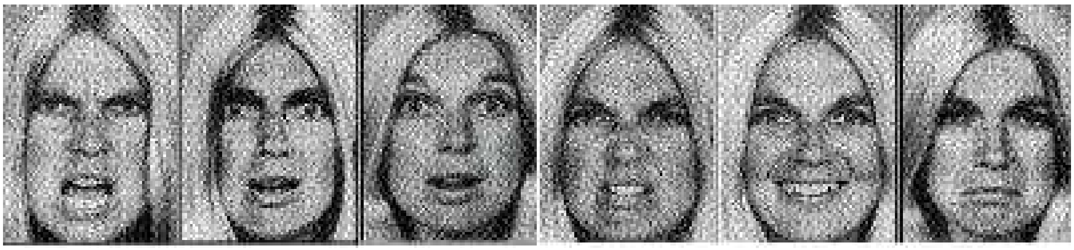


Figure 3: The 6 emotions (*Ekman et al. 1971*).

The actual purpose of nuance from these psychological perspectives is that it effectively serves to regulate the most expressive modality of human communication. The psychologist Albert Mehrabian established from his research [54] that only about 7 percent of the emotional meaning of a message is communicated through explicit verbal channels, with 38 percent communicated by para-language - the use of the voice (pitch, rate, volume, fillers). The remaining 55 percent comes directly from non-verbal modes - gesture, posture, facial expression, etc. Thus it is more often behaviour, other than spoken or written communication, that creates or represents meaning. As such, nuance does indeed seem to represent an important factor in the underlying complicity between human generation and perception that serves to regulate interaction.

While this psychological framework has arisen from the observation of the subtleties in human dynamics and how they are perceived, a significant amount of research into understanding those aspects which constitute realistic behaviour were discovered in order to create *animation*. This, after all, is concerned with the synthesis of believable form and motion in order to create entertaining and engaging characters that act and move in ways that fool audiences into accepting them. This was mainly performed by the pioneering work of Walt Disney Studios in the 1930's, when the first systematic attempt was made to identify and isolate those elements that contribute to successful characterisation. This resulted in the perfection of a number of procedures that effectively became enshrined as the fundamental principles of animation. These include the importance of such heuristics to govern: *Squash and Stretch*, *Timing*, *Anticipation*, *Follow Through and Overlapping Action*, *Staging*, *Straight Ahead Action and Pose-To-Pose Action*, *Slow In and Out*, *Arcs*, *Exaggeration*, *Secondary Action*, and *Appeal*. Factors such as "Exaggeration" are especially important for effectively relaying emotional characterisation, while other rules guarantee retaining the realism of deformation by preserving volume - as dictated by "Squash and Stretch" (as shown by the famous sketch in Figure 4 that even the simplest form can be given a subtlety of expression).

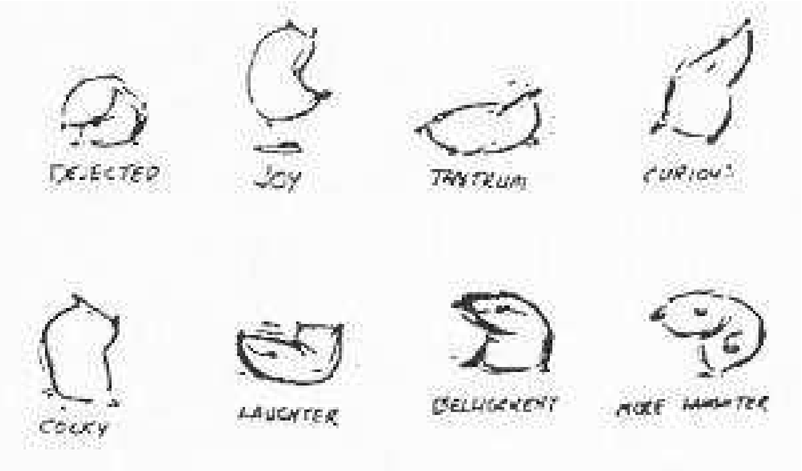


Figure 4: The Disney Flour Sack (*Thomas & Johnstone 1981*).

The first effective application of these can be seen in “*Snow White and The Seven Dwarfs*” (1937) in which the power of the characters lies completely in the control of the animation. Particular emphasis is paid to the fact that certain features of the face have a greater role to play in expressing emotion (eyes, mouth, eyebrows, eyelids), while others can be more coarsely implemented (such as lip synchronisation with actual sound). Other rules guarantee that elements must always be kept moving - never static - and that such motion must always be non-symmetric, so to avoid the problem of “twins” (e.g. eyes blinking together). In total, these ideals are listed initially in the book by ex-Disney employees Thomas and Johnston [73], but are re-iterated by Lasseter [46] due to their continuing applicability in modern computer animation. The advent of increasing computing power may have led to increasingly sophisticated imagery, but the characterisation and essence of dynamics conveyed by the skill of the animator, must still be effectively employed - as seen with Pixar Studios “*Luxo Jr.*”(1986) and by their first full-length feature "*Toy Story*"(1995). These attempts were however reliant on rigid-body models for simplicity of construction and rendering, and it is only recently that the power and flexibility of modelling the dynamics of non-rigid objects been possible in such films as “*Final Fantasy: The Spirits Within*”(2001) , “*The Matrix Reloated*” (2003), and “*The Return of the King*” (2003). This has resulted in the creation of incredibly complex and expressive characters - such as Gollum (Figure 5) whose underlying mesh of 2600 polygons (mostly quads) and 964 control points require a wide spectrum of new techniques for even experienced animators to control successfully.

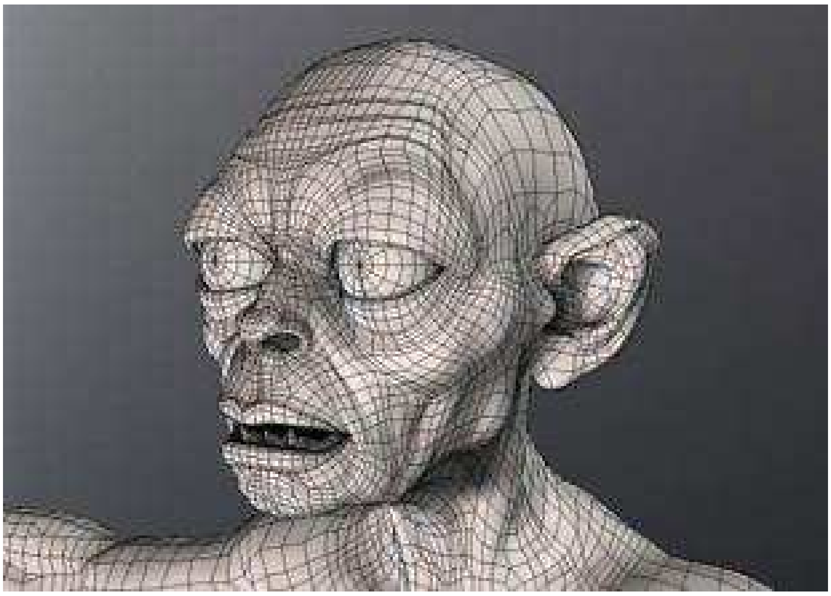


Figure 5: Gollum (*Weta Digital 2003*).

2.2 Describing Nuance

While the focus of research into understanding the perception of nuance often aims to interpret the underlying, subconscious purposes behind human dynamics - it is increasingly caught up in the problem of successfully recording the entire range and subtlety capable of being generated. That is, the possibilities for observing (either manually or automatically) and then *describing* those variations that define different interpretations, through various forms of representations and abstractions. The oldest solution with a long pedigree of transcribing the motion of humans in extensive detail, is notational systems for dance. Foremost of which is Labanotation, first published in 1928 by Rudolf Laban and continually developed and employed by many practitioners for describing a wide range of styles and disciplines. This scheme is reliant on a very simple motif based form capable of recording across time for both general movement (allowing for creativity in its exact interpretation), or the exact specifics of a particular movement (e.g. a "Spanish dance" hand gesture in which the fingers, one by one, perform a delicate folding, gathering action leading with the little finger, while the arm, moving in space, rotates outward). The basic scheme has been used to successfully create a large number of "scores" for every conceivable form of dance and expressive movement, and has even been converted from motion-capture data by Matsumuto *et al.* [53]. However, it is more a relative, as opposed to an absolute, mode of directing movement - like a musical score, in which the role of a choreographer (like a conductor) must provide the finer points of interpretation (Figure 6).

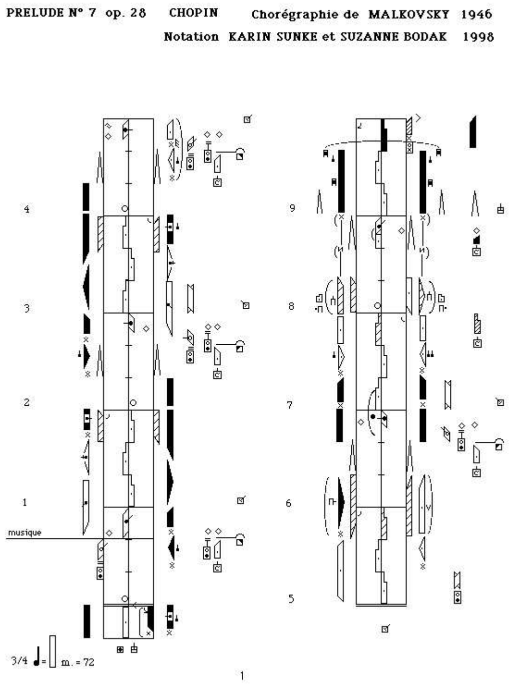

PRELUDE N° 7 op. 28 CHOPIN Chorégraphie de MALKOVSKY 1946
Notation KARIN SUNKE et SUZANNE BODAK 1998
4
3
2
1
musique
3/4 ♩= m. = 72
1
9
8
7
6
5

Figure 6: Example labanotation score (*Sunke & Bodak 1998*).

A more recent approach is driven by this need for absolute control in the problems in generating virtual "Avatars" and "Social Agents" for improved human-computer interaction and realism in virtual environments. The desire is again to be able to quantify and represent the possibilities of facial and body dynamics in order to offer control of animated sequences to higher-level procedural control. This can include aspects of facial expression and gesture to be synchronised with speech performed by Cassell *et al.* [21, 20], or for more general scripting of behaviour for virtual actors as advocated by Perlin [62, 63]. Such models thus represent attempts to directly encode theories from dialogue based behavioural analysis and social psychology - for example, Nakata directly applied somatic theory to recreate expressive movement [56]. Consequently, they naturally represent control in terms of the abstraction they are to support - constructed from hand-coded basic atomic sequences and symbolic rules that can be mapped and combined to form desired behaviours on cue. The fact that such models are not trained or based on actual human data also leaves them susceptible to the flaws and difficulties associated with traditional animation, and with the reality that they will thus inevitably omit a higher level of accuracy and realism.

Such work has also led - in combination with the recent popularity of *markup-languages* - to a new set of specifications for standardised and transferable descriptions of human dynamics within an XML framework - such as the Virtual Human Markup Language (VHML) and the Human Markup Language (HuML). These allow quite specific, high-level descriptions of the configuration of the body, and how it then moves in response to time and circumstance - but still neglecting to describe any exact shape change or precise positioning, in order to allow control to generalise to a whole range of models. The ultimate in current technology for encoding such "audio-visual" objects is the MPEG-4 standard, a significant part of which is taken up by the Synthetic/Natural Hybrid Coding (SNHC) that aims to directly address the definition of virtual human faces. This is divided into two main components - Facial Definition Parameters (FDP's) for customising a baseline model to a suitable degree of realism, and Facial Action Parameters (FAP's) to then control the animation - following studies into muscle action units to define 68 groups of expressions and visimes organised into 10 groups corresponding to regions of the face. An example based on this is the work of Byun *et al.* [17]. They consider an already well refined symbolic representation for expressions of the face (FAPS), on top of which they map values from the "effort" parameters associated with Laban Movement Analysis - dimensions such as: space, weight, time and flow (Figure 7). The value associated with each parameter acts as a filter to perturb each FAP for each frame of motion, and the combined effect can result in a broad range of subtly different synthetic expressions. This research represents a more top-down approach to the problem - in that the abstractions of the face are predefined and directly controlled along the lines of more traditional animation techniques, and consequently lack a foundation based on actual data based observations.

This standard is actually derived from the more fundamental Facial Action Coding System (FACS) [31] originally devised by Ekman and Friesen to provide a more fine-grained recording scheme for expressions. This serves to divide the face into upper and lower regions - with 46 associated *Action*

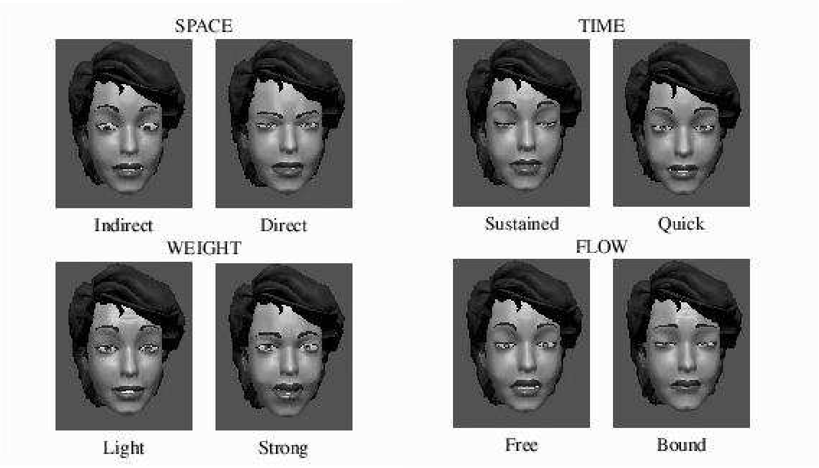


Figure 7: FacEMOTE (*Byun et al. 2002*).

Units (AU's) which are further structured into vertical, horizontal, orbital and oblique movements that are directly correlated to the muscles of the face. Most importantly, it systematically breaks-down and defines the possible combinations of Action Units - including the extent to which subtle differences in the modulations of the actions create completely different results (Figure 8). FACS therefore represents the most comprehensive scheme for describing upwards of 3000 combinations for different facial expressions, and does so with regard for nuance, by defining the fine distinctions between combinations of expressions that are particularly susceptible to miss-classification and ambiguity. Such distinctions are however quite subjective - such as: "*In 43E alone the eyelids appear relaxed. In 7+4E3 the eyelids are tightened together, not relaxed.*" - and despite the system being recently revised, it still has a significant problem in successfully applying it to studies as it takes upwards of 100 hours of training to successfully attain a reasonable level of competence in isolating and scoring individual performance. Some approaches have attempted to construct expert systems to assist with the automation of this task (e.g. Pantic *et al.* [60]) but are however still reliant on manual identification and direct measurements of features in input images of faces.

This lack of exactness and reliance on human assessment forms a common factor of all the approaches described above. Another issue is the fact that human expression is *multi-channel* - that is, at any one instance a number of components may be active (the eyes, mouth, hands) which make it hard for an observer to directly focus and recognise exactly what is occurring in combination. These problems have understandably led to a desire to automatically acquire the complex dynamics of the human form - particularly in order to form a classification for the observed expression, gesture, or individual. In particular, a number of approaches have been developed to analyse images for the automatic detection of facial actions - a excellent survey of which is presented by Bartlett *et al.* [7]. These can be summarised by the underlying means with which the data is pre-processed in order to

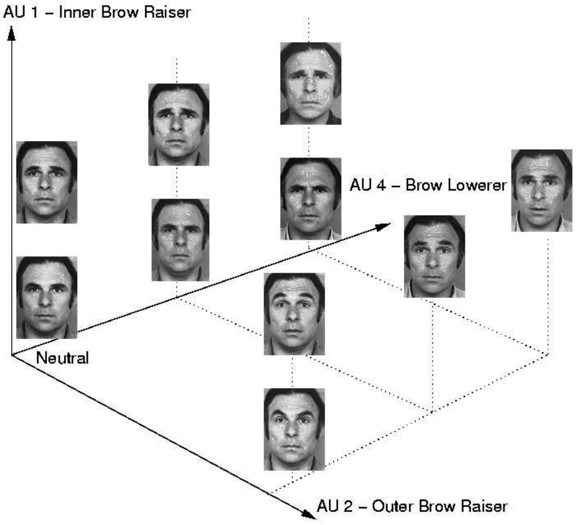


Figure 8: Four FACS Action Units and their combinations (*Ekman & Friesen 1978*)

determine the best intrinsic representation for basing a classification. Such distinctions also extend to other approaches that seek to tackle motion for the whole body, and call for a wide spectrum of Computer Vision techniques in order to first accurately capture the data that embodies nuance. More details of these specific techniques can be found in any good vision reference textbook such as [34], and can be generally grouped as follows:

Motion-based descriptions. Exploit *optic-flow* to provide detailed fields describing the variation in surface movement. For example Yacoob and Davis [84] based their work on theory from psychophysics proposed by Bassili that a certain minimal spatial arrangement of features are sufficient for identifying expressions. Based on this idea they created mid-level template representations based on the optic-flow output calculated for localised and tracked face regions (e.g. mouth, nose, eyes) - in order to establish a classification of one of the 6 original Ekman expressions. They later extended this model for complete spatial-temporal training of a radial function based neural-network on different expressions. Further enhancements and improvements were performed by Hoey and Little [41] in applying higher-order Zernike polynomials to better represent the intrinsic complexities of non-linear motion (comparative to PCA based holistic approaches). An idea of the shape can also be interpreted from the raw motion to yield a factorisation able to recover a shape matrix from image sequences, as in Jebara *et al.* [42] to directly construct approximate 3D models of the head. Similarly Brand [13] recreates the complex structure of the face from the observation that both non-rigid 3D structure-from-motion and 2D optic flow can be formulated as tensor factorisation problems. Thus, they can be made equivalent, such that it is possible to simultaneously track a non-rigid surface and acquire its shape basis simply by manipulating the rank of the flow calculations. Recovering more accurate motion and surface deformation is becoming more possible with the advent of recent advances in both stereo and range-capture which overcome their inherently static nature, enabling *real-time* (or at least off-line) range acquisition over complete temporal sequences of a few seconds duration. In particular, the calculation of *range flow* [68] analysis serves to define the velocity field that describes the motion of a deformable surface, by extension of similar calculations for 2D optic flow. The extra information that this provides can in turn be exploited in order to resolve temporal non-rigid motion and construct models of deformation for the human body, for example Nebel and Sibiryakov [58] (Figure 9).

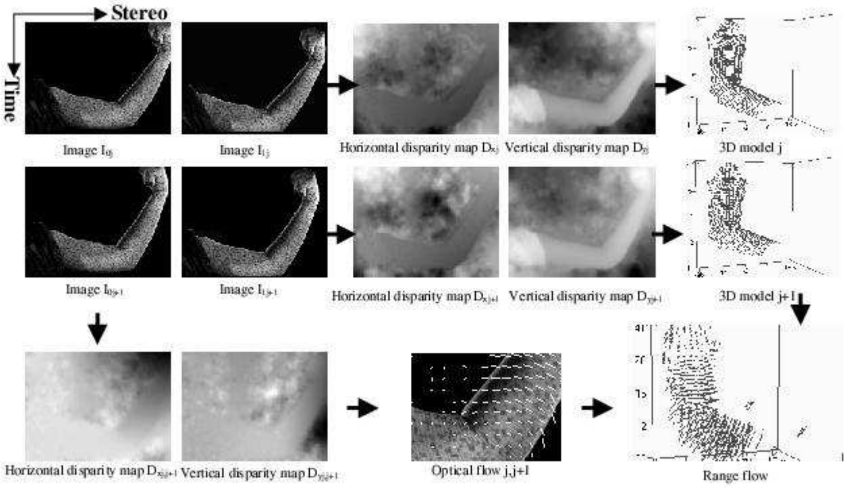

Stereo
Time
Image I_{0j}
Image I_{1j}
Horizontal disparity map D_{xj}
Vertical disparity map D_{yj}
3D model j
Image I_{0j+1}
Image I_{1j+1}
Horizontal disparity map D_{xj+1}
Vertical disparity map D_{yj+1}
3D model j+1
Horizontal disparity map D_{xjj+1}
Vertical disparity map D_{yjj+1}
Optical flow j,j+1
Range flow

Figure 9: Calculating range-flow for the arm (*Nebel & Sibiryakov 2002*).

Model-based descriptions. Employ *explicit physical model*s to which the data is fitted. For example Essa and Pentland [32] extended the original work by Terzopoulos and Waters [72, 80] to create a very detailed anatomically correct face model - including 44 muscles and elastically deformable skin. Input from actual faces were then mapped to the model via a number of common feature points. Classification could then be made by analysing the parameters of the model, on the assumption that these accurately reflected the reality of how people form expressions and that the mapping preserved these distinctions. One solution to this issue of accuracy is to base the construction of the model on a number of discrete range-data sample samples across a range of movement - for example the use of subdivision surfaces based techniques to encapsulate the different scales of movement of the body surface as performed by Allen *et al.* [2], or the creation of volumetric representations for the physical properties of soft tissue by Nebel [57]. Such techniques are moreover tied up with the specific issues encountered for accurate construction and representation of humans models for Computer Graphics from range-data. The issues with this are presented in [40, 52], and highlight the problems of capturing and representing the complexity of natural surfaces, from often incomplete and noisy data.

Feature-based descriptions. Rely directly on a *set of feature*s extracted from an image. For example, the recent application of Active Appearance Models (AAM) developed by Cootes [24] provides a robust set of tracked points on the basis of shape and texture. These can be used as input vectors to a variable length Hidden Markov Model in order to train a set of spatial-temporal trajectories that represent different components of expressions as described by Bettinger, Taylor and Hack

[8, 38]. Such a model can then be used to successfully synthesise a range of realistically novel expressions and behaviours, although none are explicitly defined as a particular expressions, but instead form a set of similarly clustered configurations (Figure 10). This is in turn part of the general problem in being able to "track" human motion via purely *passive* means (devoid of any artificial features) directly from images of people - a comprehensive survey of which is presented in [55, 1]. Such an idea can also be applied to describing full body motion, but requiring a solution to the combinatorial problems of multiple degrees for freedom in order to fit articulated models to observation, for example using particle filtering by Deutsher *et al.* [28] or point correspondences by Taylor [70].

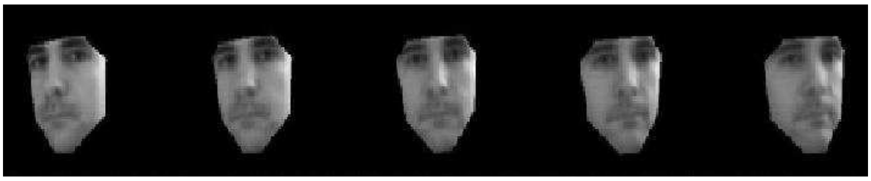


Figure 10: Synthesised head behaviour trained from an AAM (*Hack et al. 2003*)

Holistic-based descriptions. Apply *statistical techniques* to complete images in order to discover intrinsic dimensionality. For example, Choudhury and Pentland [22] built motion field histograms from a sequence of faces, to which PCA is applied in order to extract the top eigenvectors. These form a feature set that is able to model the pertinent temporal components of various expressions - which are relatively robust to rotational and translational errors. Alternatively, template matching approaches, such as Bobick and Davis [11], provide another means to selectively build a classifier able to discover relevant features - while overcoming issues with the "alignment problem" between individual images. Such models can also be used to learn correspondence to external control signals, such as an audio track, in order to to predict the dynamics of one from the other as proposed by Bregler [16, 15] and extended by Brand [12]. These seek to model the underlying face behavioural manifold - a surface of all the possible facial pose and velocity configurations embedded in a high-dimensional measurement space. Approximating this space and the transitions that occur within it as a probabilistic finite state machine (a HMM) can allow it to describe the most likely state given a signal - creating a rich and plausible description of how the face realistically alters in response to speech. In describing the actual "purposefulness" of full-body human motion, the work of Wren *et al.* [83] sought to discover the effectual alphabet of behavioural combinations as a probabilistic model constructed from low-level blob features. Alternative approaches have investigated the training of Artificial Neural Networks for the task of recognising facial expressions by Lisetti and Rumelhart [51], or facial emotions by Dailey *et al.* [25]. These have shown that only particular areas of the face and relatively simple networks are required for basic recognition, but do require careful regard for network topology and input selection.

2.3 Interpreting Nuance

Besides seeking a means of robustly identifying and describing the basic forms of expression that can occur, there is also the more fundamental issue of *interpretation*. That is, extracting and modelling the true underlying nature and invariant properties of human dynamics. As stated in the introduction, we define “nuance” as the *variations in dynamic shape and motion that serve to distinguish acts of a similar basic form*. This is a unifying statement that can also be used to describe a number of key works that, while not using the exact term of “nuance”, share commonality by seeking to establish the distinction between the variations of *form* acting on top of a basis of common *structure* that together define expression. As such, we consider this a key concept that (more than any other) defines our specific field of research, focusing on the separation of *style* from *content* - an idea that is most succinctly expressed by the core paper of Tenenbaum and Freeman [71]. In this they define the open question as how to explain the human perceptual ability to freely separate these two factors - a crucial task that allows to successfully generalise to novel and varied forms, to anticipate the style of missing elements, and to distinguish and appreciate the fine degrees of subtlety in various modes of expression. They give examples of how this can be applied to handwriting, accented speech, and faces under varied lighting conditions - in each case showing how a straightforward bi-linear model can learn approximations across the effective two dimensions of style and content. Their conclusions were that such a model offers a simple and general framework for how biological and artificial perceptual system can solve a wide range of *decoupling* tasks given suitable input representations.

From this understanding, a number of approaches can be seen as attempts to further tackle this issue of distinguishing between the components of any perceived act (including the temporal aspect). This is especially evident within the domain of human dynamics, which form a good example of how similar base forms can be modulated into differing perceived interpretations. Such research is driven primarily from a Computer Graphics desire to provide synthesised control for animation based on motion-capture input from humans performing exemplar motions. The problem is that such data only represents finite fixed training sets for “set-piece” events. In order to be truly reusable (and useful) there is a need to support *motion-editing* in order to derive the underlying nature of the dynamics - to alter initial positioning, to blend and stitch across novel transitions, and to re-target them to models that vary in dimensions to the original. In Computer Vision, the ability to separate form and structure would also provide a means of greatly simplifying the problem of classification, by effectively decoupling the differing modalities in any given movement and making them invariant to individual instances and viewpoints. Works that attempt this extrapolation are based on a number of common themes, and can be further grouped by their similarities in applying various techniques to the problem as follows:

The application of **frequency analysis** techniques to distinguish low (content) and high (style) frequency modulations, form the basis of a number of early approaches to the problem. This is based on the known effects of additive noise to computer generated imagery and motion in order create a greater semblance of realism (as first described by Perlin [61] and illustrated in Figure 11). A key paper is that of Unuma *et al.* [75] who were the first to be able to transfer the emotional context of different locomotion style in order to create novel superpositions via Fourier series expansions. This was reliant on periodic example motions, in which the connection to a kinematic model was maintained (e.g. step spacing) and from which the high-level characteristics of particular gaits could be extracted and reapplied. Similarly, the work of Lee and Shin [47] adopted a multi-resolution approach in which the motion capture data was represented as a collection of coefficients from coarse (global pattern) to fine (detail patterns). By exploiting these they were able to attenuate and enhance the fine level detail in order to vary the effect and impact of motion events. Both approaches represent the fundamentals of how movement in humans exhibit repeatable and unique properties, a factor exploited in recent biometric research into identification based on walking pattern. They are however, by their very nature, constrained to particularly periodic wave forms, and are consequently limited in only being able to reapply their effects to the original or intrinsically similar motions - making no distinction as to the similarities and variations between samples, or which frequencies or subdivisions truly encapsulate any given “emotion”.

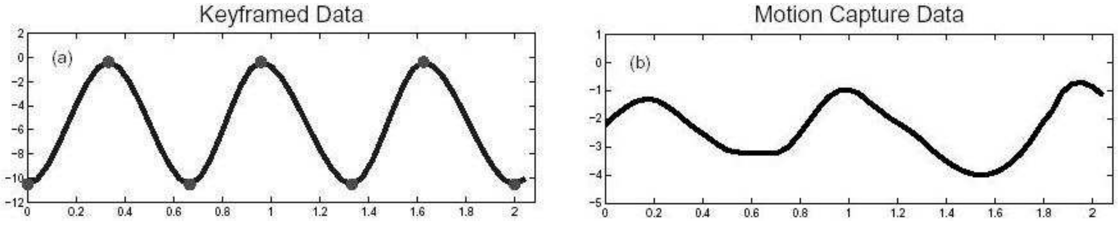


Figure 11: Contrasting synthetic key-frame and real data (*Pullen 2002*).

Improvements to this have been inspired by recent ideas about texture synthesis, leading to the concept of a **motion texture** that can act to divide data directly into its dynamic and stochastic elements. This term was originally coined by Pullen and Bregler [65] in their work that allows animators to construct a sketch of desired motion with a limited set of key-frames, and to then apply texture from example sequences of motion capture of a particular style. The key to this was in being able to match and “fill-in” the motion curves for all joint positions by progressively searching from low to high frequency bands in the example data - enabling novel sequences to be generated in the style of whatever motion-capture examples were searched. Li *et al.* [49] define texture more rigidly as the set of motion “textons” and their distribution. Specifically, a texton is modelled by a linear dynamic system incorporating low-level detail, and a distribution that is represented as a transition matrix recording how likely one texton is to switch to another - defining high-level choreography. Thus, entirely new styles can be synthesised by applying noise to the textons, while retaining the underlying form dictated by the transitions. However, this noise is undirected in its desired effects (i.e. can result in physically implausible dynamics), and attempts to radically edit the distributions for completely different sequences do not work well. Again, these approaches therefore work best for periodic actions at a particular frequency, and for re-targeting to sequences with similar constraints.

An alternative approach seeks to directly derive **action parameterisations** that distinguish atomic modules for acts and how they can be combined and controlled. Rose *et al*. [67] advocate discovery of motion "verbs" and their controllers as "adverbs" which represent emotional axes (such as happy-sad) or physical possibilities (such as upwards-downwards). They achieve this by creating a continuous verb interpolation space, where the dimensionality equals the number of adverbs, created from the set of example motions modelled by a combination of radial basis function and constrained by a kinematic model (Figure 12). An interpolated point in this space then represents any given motion for a combination of affecting adverbs, and defines the motion characteristics at that moment. This work was in turn extended by Jenkins and Matarć [43] to build a complete repertoire of action and behaviour primitives for modularised humanoid robot control - focused on actually discovering the "verbs" themselves by applying non-linear dimensionality reduction to the raw data via an extended variant of the Isomap algorithm capable of finding spatial-temporal structure. This resulted in a number of behaviour primitives being discovered, in turn controlled to varying degrees by whatever parameterised action primitives are used. Both these works thus represent a supervised learning approach to symbolic discovery of underlying motion and parameterisations - but suffer from limitations in that the initial data must have very clear distinctions for specific gross movement, and so do not accommodate much by way of subtlety.

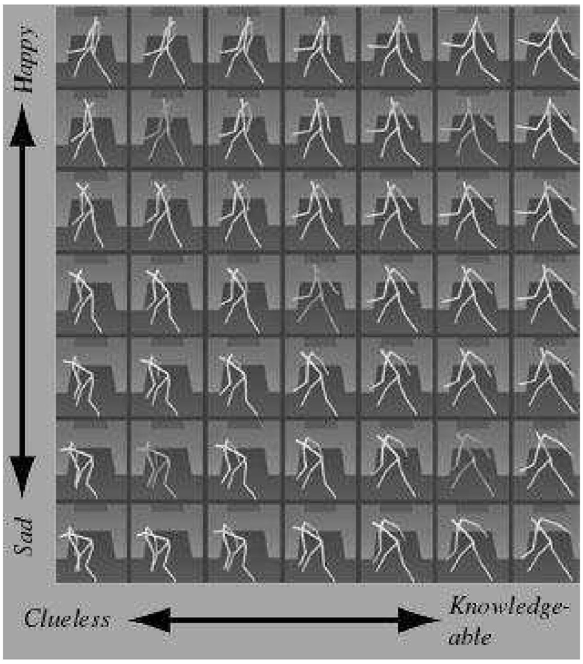


Figure 12: Verbs and Adverbs (*Rose 1998*).

Further research has extended to a more statistical based form of **HMM decoupling** for style and structure based on direct analysis of high-dimensional spaces - with no preconceptions of what "verbs" or parameterisations may be present. Brand and Hertzmann [14] achieve this in the form of a "stylistic" Hidden Markov Machine - casting the problem as a pure unsupervised learning problem on motion capture-data, by which they desire to capture the data's essential structure after removing any particulars of individual samples. This can then be further separated into those states that represent the basic transitions of the motion dynamics, along with an independent stylistic parameter that produces variation to the output of each state, all learnt through a process of EM based optimisation. The model created was even able to distinguish the subtle distinctions between the style of an expert and a a novice performer, and further analysis across all distributions enabled discovery of a parameter subspace incorporating various stylistic "degrees of freedom" that can be exploited to infer entirely new, unseen, variations (Figure 13). An alternative approach, involving a number of HMM's coupled together with an expression intensity component, was proposed by Lien et al. [50]. This was directly applied to facial expressions based on the Facial Action Coding System - with video input analysed along a number of feature points, optic flow, and high gradients (e.g. brow furrowing) in order to build the training vectors. Data was only acquired for frontal video under well lit conditions, and was further reduced by PCA before being used to train the system to recognise various expressions - where intensity is measured by the distance in the resulting eigenspace to similar motion correlations. Ultimately, the HMM (one for each action unit) returning the maximum output probability is then accepted as the classification for any novel input, and the degree of combination of AU's could be seen by the response across all classifications. This approach is thus much simpler than the more holistic approach adopted by Brand, which is more complex in its ability to extrapolate the variation which embodies "nuance" - although, for larger scale motion and relatively limited set of styles. Furthermore, both approaches are, as with all data-driven methods, reliable only to the limit of their training instances.

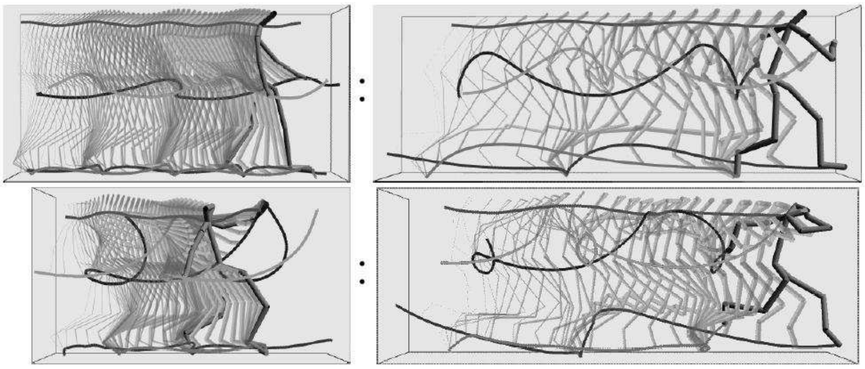


Figure 13: Walk, Run, Strut and synthesised style (*Brand 2000*).

Other works seek a means of directly discovering the distinctions between form and content, in line with the original proposal of Tennbaum, by focusing around the idea of **n-mode decomposition**. This extends from the initial work of Vasilescu *et al.* in developing "Tensorfaces" for multi-linear analysis of natural images [77], and extended to the extraction of temporal "motion signatures" of an individual's distinctive pattern of movement and invariant style applied to the action [78]. This is performed through a SVD decomposition of high-dimensional tensor representations for human motion, to derive orthographic matrices for the variation and co-variance of people and actions (the “modes”). Novel superpositions can be then achieved for new people or actions by straightforward tensor flattening and multiplication. Wang and Ahuja [79] also use a Higher Order SVD to discover separate expressive and facial subspaces invariant to each other. These were constructed from a corpus of cropped images, and further reduced in dimensionality by a fitted deformable model, from which their algorithm decomposed the resulting tensor of person, feature and expression into orthogonal basis sets. A similar idea to these approaches, but using a more structured composition of 3-modes (pose, style and time) is performed by Davis and Gao [27]. Again, the idea is that a decomposition provides explicit separation into a low-dimensional set of components, from which tunable weights can be applied to in order to control stylistic trajectories. Such reduction is similar to traditional PCA, but instead representing basis sets as a 3D cube rather than 2D matrix and thus modelling multiple axes of variance across a dataset as opposed to one. Such styles can be expressive or physical modes provided by exemplar motion-capture (male-female, heavy-light, fast-slow), and the system learns from the a set of weights by gradient decent, in order that emphasise those trajectories most indicative of the distinct modes. Chung *et al.* [23] directly implemented learning of a bi-linear model of facial expressions, carried out on texture and features derived from active appearance model processing of video (i.e. shape tracking of mouth, eyes, jaw line, etc.). Again, PCA is applied to reduce the dimensionality of the data, from which training vectors are factorised into either symmetric (style decoupled) and asymmetric (style incorporated) models. SVD decomposition of these models allow generalisation for style and content, being possible to apply them to novel sequences in which the speech shape of the mouth (content) is maintained independent of the expression (style) of the overall face (Figure 14). These approaches thus represent a common technique that has served to successfully construct a model that can encompass a given number of distinct variations as represented by the diversity of the data set - but limited to an initial degree of selection for suitable sets of features, and an organisation into whatever modes are deemed to be present.

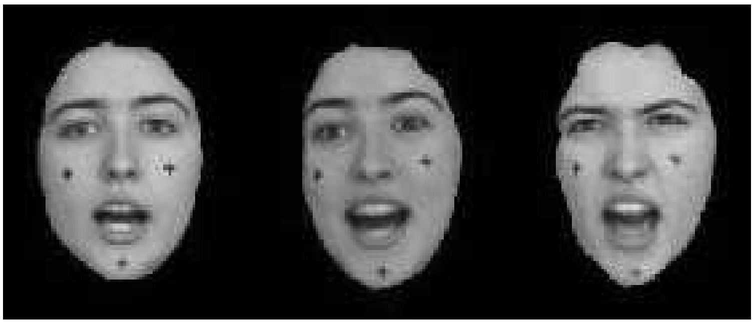


Figure 14: Neutral, happy and angry generated from the same face (*Chuang et al. 2002*).

2.4 Critiquing Nuance

In summary: all of this prior work embodies an insight into nuance. Though not directly calling it as such, there is a common idea and justification for the ability to generate and detect the more subtle variations that define activities and expressions. Humans have this ability built-in as encodings in the brain that can cope with an almost complete surfeit of information - as shown by psychophysics. Less rigorously defined, but just as applicable are the results of the research into more behavioural formations of expression - in particular the perception of emotions in the face and how these vary to provide us with the information to establish the degree that someone is “confused”, “sad” or “happy”. While we have described and highlighted a number of issues with the various schemes, approaches and techniques for dealing with nuance, their are a number of key points that can be reiterated as follows:

- **Representation**. Much of the work into how to represent of nuance can be viewed as effectively the problem identified by Badler [6] that “*what is needed is an approach to the representation of human movement which accepts the diversity of information sources and yet provides conceptually tractable data and control structures for the expression of movement*”. Thus, all descriptions of movement are reliant on instances of “primitive movements” to provide the basis set for all expression. The problems with symbolic techniques such as Labanotation is that they then require additional expert interpretation in order to successfully “inject” expressiveness. Similarly, the behavioural encoding systems used to transport Virtual Character descriptions (VMHL, MPEG-4 SNHC), and those procedural languages for controlling animation, ultimately require hard-coded “engines” to interpret and generate natural looking motion (based on the skill of the animator/programmer). The Facial Action Coding System supports a more varied and descriptive scheme that has been rigorously defined from actual observation, and which directly tackles the issue of differentiating ambiguous expressions by describing subtleties. It does this with further regard for the fact that the face represents the most compact and visually diverse source of human dynamics, and yet also varies from person to person by the greatest degree. This diversity makes the recognition of all distinct Action Units a hard problem even for humans - which has effectively limited FACS in its application. However, it continues to form the representation which can most meaningfully correlate nuance to expression, creating a meaningful framework to base research on expression.
- **Precision**. The temporal and spacial characteristics of dynamics are what ultimately define nuance - and it is these aspects that must be precisely captured. Thus, the problem is essentially one of gaining as much data as possible describing the structure and motion characteristics, which may then reveal upon further analysis those components that best describe the sequence in terms of those elements of nuance. While motion-capture and other feature-based tracking techniques have been used to great effect for extracting exemplar movement, they have failings in their level of accuracy and in their complete absence of shape information. This in turn

raises the more fundamental issue of whether the dynamics that are captured (being only a sample) are truly representative of whatever sequence is being recorded - since they provide only what is predetermined by sensor placement or the nature of the feature detector used. The disadvantages are also in the fact that the information captured is only valid for discrete events and include the potential for alteration in marker positions due to soft-tissue movement and tracking problems through deformation and occlusion. Similarly, model-based techniques can fail from errors in initial fittings to observations, and by the overburdening complexity of then realistically handling natural motion. Most success in capturing the more subtle and distinctive elements of expression are based on motion and holistic methods that are able to treat the rich 2D domain of images, but are thus limited to accurate 3D recovery. Consequently, the advent of temporal range-scans offer a new direction that is able to combine the accuracy and richness of shape definition with interpretation by extended flow and statistical techniques for discovering the intrinsic dimensionality of expression.

- **Seperation**. The analysis of nuance is implicitly tied up with the ability to separate the features which define the structure of underlying motion, from their variations that define the form of different expressions. As such, dynamics can be seen to occur over an effectual frequency of ranges - from the coarse scale of gross movement, to the subtle scale of individual style. Such a view is traditionally limited to relatively large-scale cyclic actions such as walking style and dancing steps - and further decomposition is therefore required in order to break-down a sequence into individual “verbs/adverbs” or “modes”. Most of this research is in turn focused on a generally small subset of expression (joy, fear, fast, slow, etc.), or is reliant on human intervention to define regions of interest. This is mainly due to the high-dimensional spaces that must be processed - requiring novel statistical techniques to differentiate the axis along which variations occur. A much richer initial set of features (other than existing approaches based on motion-capture) pose even greater challenges for successful interpretation - placing the onus on other techniques able to construct truly representative temporal feature sets. Other unsupervised techniques may thus yield an approach that is better able to cater for higher dimensionality, and can act as input to better models for the classification and synthesis of expression.

This work seeks to address a solution to these main points, and thus contribute to an improved understanding via nuance and its role. In particular, no prior work (of which we are aware) combines high-resolution 3D capture of human data with statistical analysis that can identify a truly representative feature set of the dynamics of expression and how these can be further decoupled into variations over a common form. By furthermore correlating the resulting model to an established psychological framework - the Facial Action Coding System - we would be able to directly interpret the validity of nuance, and to hopefully apply it as the basis for classification and synthesis of realistic expression.

3 Approach

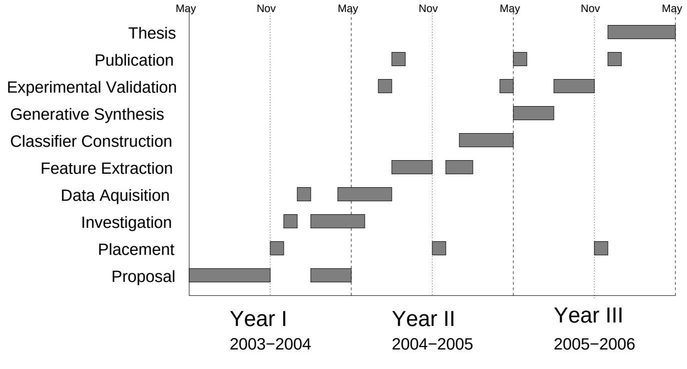
In this chapter we detail the proposed approach to solve the problem defined in chapter 1, and which builds on the prior background of chapter 2. This can be best achieved by division of the work into various stages designed to meet our key objectives as follows:

- Scientifically show that the hypothesis of *nuance* embodies the variation in dynamic features that define realistic human expressions.
- Provide techniques for successfully creating and segmenting range-flow captured from stereo, coupled with enhanced statistical methods modelling the complexity of human dynamics.
- Apply the resulting technology in order to provide Virtual Clones with a way to capture and automatically create a range of expressions for captured heads (to "bringing life to animation").

We seek to do this by focusing on the following open questions that have been raised by the initial hypothesis, and have also emerged during the course of our discussion in the previous section. These provide the focus to our ultimate aim of experimentally validating the concept of nuance as a principle that can be applied to generate realistic and recognisable human expressive modes.

1. **"Is a model incorporating such nuances more realistic than those without?"**
2. **"What role do particular nuances have in distinguishing expressive modes?"**
3. **"To what extent can nuances accurately reflect the variations of human expression?"**

For the overall structure of our approach, we follow the path of a "classic" machine learning methodology as shown by the time-line figure below:



3.1 Investigation

Goal: to validate our initial concept of how nuance can be modelled, and to learn to exploit the technologies available to meet our requirements for its capture and analysis.

Before embarking on the main body of work, we seek to confirm the presence of nuance as we have defined it in the introduction. That is, we wish to conduct some initial experiments to verify the extent to which even the simplest gesture can be nuanced as a result of adapting variations along key aspects (such as timing and direction). We are furthermore interested in gaining insight into what modelling techniques and technologies can provide us with the necessary framework with which to acquire the data reliably. This requires us to build some prototype systems in order to assess the extent and effort that may be required in order to achieve accurate results.

One such task is based on the services provided by EdVEC (the Edinburgh Virtual Environment Centre) using the Motion Analysis 8 camera optical capture system (`www.motionanalysis.com`) to obtain data for instances of simple pointing and beckoning gestures - using only markers for the tip of the right index finger and wrist. This will create a very simple dataset, yet one that hopefully provides us with one the most basic cases of a nuanced activity. Investigating the ways in which the trajectory of this single point can be modelled, segmented, classified and synthesised will be the first step to understanding nuance for the case of one "feature" - and providing the necessary foundations for the later work. In particular, we shall look at how the dimensionality of the point as it moves can be reduced - by 4D spline fitting, PCA, and probabilistic methods.

Another task involves us constructing a stereo capture system using software and advice from our sponsors Virtual Clones. With this we wish to streamline and automate as much as possible the system of capturing 3D head models - especially in terms of removing background noise and for capturing a complete temporal sequence using "burst" mode camera functionality. While doing this, we also desire to establish the accuracy of the data, by comparatively assessing the error between a laser-scanned test object and the same object captured via stereo - exploiting the idea behind the *Iterative Closest Point* algorithm to directly compute the fitting between them.

In doing these tasks, we shall also gain familiarity with various software toolkits, such as:

- GNU Scientific Library "*a numerical library for C and C++ programmers*"
- OpenCV "*algorithms and sample code for various computer vision problems*"
- OpenGL "*The industry's foundation for high performance graphics*"
- GTS Library "*functions to deal with 3D surfaces meshed with interconnected triangles*"

An appreciation gained during this stage for what is possible and tractable using current technology will serve as an important "reality-check" for the remaining duration of this research.

3.2 Data Acquisition.

Goal: to acquire highly accurate 3D spatial-temporal data of people making a range of expressions encompassing a set of nuances.

Our intention here is to build a relatively large corpus of training (and test) data sets that encompass the versatility of the human face. Exploiting the face provides us with a compact and diverse source, which forms the primary focus of human non-vocal behaviour. Unlike previous work, we do not wish to base any assumption on the initial selection of features - instead we wish for a complete 3D dataset that reflect a number of different instances, performed by people with varied different facial characteristics. To this end, the Facial Action Coding System [31] provides a suitable framework for determining our expressive modes - with its 40+ Action Units identifying up to 7000 different facial expressions in combination.

The "captures" we wish to perform are those which we believe will incorporate a suitable variety of nuance. Starting in each case from a "neutral, straight ahead" position - we will ask 20 subjects (ideally selected from range of different people of different backgrounds and ages, with half male half female) to perform 5 instances of the following general expressive modes:

- **"Surprise"** - e.g. Brows raise, mouth opens, eyes widen.
- **"Happy** - e.g. Brows knot, mouth opens, eyes narrow.
- **"Angry"** - e.g. Brows knot, mouth tightens, eyes narrow.

These are chosen because while they cumulatively form very different final expressions, they share commonality between component action units - a factor that is known, for subtle variations, to give rise to a certain amount of ambiguity. Furthermore, we are not concerned with any mode that incorporates gross head or eye movement, nor shall we bias the performance of the subjects - but shall simply tell them immediately prior to the capture of the mode they are to perform (and so hopefully elicit a relatively spontaneous response).

Actual capture shall be performed using a binocular-stereo digital capture rig, providing high resolution images at 3072 x 2048. We hope to exploit the "Burst Mode" of the cameras - supporting capture of 4 images in rapid succession at 2.5 fps (= *1.6 second duration in total*) for the temporal aspects. This should be enough to capture the relatively fast speed of all the expressions. Thus each capture will total (4 x 7MB raw) 28 MB of data. These shall have to be in turn post-processed by software to first perform background removal and registration so that only the head remains, and to recover 3D depth information via stereo correspondence. At an estimate each capture of 4 images will require 1 hour of processing to build. Thus, the total size of (5 instances x 20 people x 3 expressive modes=) **300 captures** will be 8.2 GB of raw data requiring at least 20 days of processing for 3D recovery. We also anticipate that this dataset will be extremely useful for other research - there being no other collection of highly accurate spatial-temporal head models of which we are aware.

3.3 Feature Extraction

Goal: to establish the nature and structure of the data and its intrinsic dimensionality with respect to nuance.

This is an issue of reduction - both temporally and spatially - performed on the datasets in order to diminish its dimensionality to those aspects that embody nuance. Up to this stage we have only suggested from observation what these may be (e.g. timing, co-ordination, etc.). Given the wealth of data acquired, we now hope to establish the existence of such underlying features.

Our initial step is to smooth and re-sample the input data into a more useful representation - while importantly retaining all the subtlety. To this end we shall calculate the mean/average neutral head position across the entire data-set and use this to align all the positions of each capture (given that we have already removed any background and upper torso points). We shall then calculate *range-flow* for every point exploiting the intensity values by the point-to-point correspondences with the original image data - as defined by the algorithm developed by Spies [68]. This will produce a data-set starting at time t_0 for their initial X,Y,Z coordinates and colour, and proceeding with the range flow field displacements and colour changes across t_{+1}, t_{+2}, t_{+3}. We may then interpolate between these key poses in order to enrich the temporal aspect to the data, and we may also divide the data into respective action unit areas such as the brow, mouth, nose, eyes, cheeks, etc. for better tractability in processing.

On the basis of this enhanced/segmented data, we intend to first visualise the results of the range-flow calculations. By simply viewing the average displacements for all points projected onto the t_0 image - we would hope to see different regions of the face exhibiting variation in velocity, changes in positions, and relative to other regions depending on the expressive mode. Mean calculations across sets of expressions would hopefully prove even more evident. Some of these may appear to be very similar, but from this we can then experiment with unsupervised techniques for segmentation and extrapolation of prototypical features and trajectories of points (e.g. k-means, clustering, SOM), and supervised (3-mode PCA [27], the Isomap algorithm of Jenkins[43]).

For example, we could find interesting feature points in the spatial-temporal sequence by an approach which computes histograms over points P^t, ..., P^{t+s}(i.e. range flow) and selecting those which exhibit maximum variance. Such a solution could then be used to segment the data by grouping neighbouring similar points into a number of regions (nearest-neighbour) - and then taking the mean centroid of each region to define the dynamics. Another possibility is to form a sequence of displacement or depth maps from the 3D data, and to then apply traditional image processing feature tracking (e.g. on Harris points).

These techniques will thus seek to effectively provide a form of “intelligent” motion capture - such that by holistically considering both form and motion, the feature set derived serves to best describe the most salient and informative points of the dynamics. Analysing the variation that then occurs in these points over different sequences of examples, would then reveal those parameters which differentiate nuance - aspects which must be modelled by the next stage in constructing a classifier.

3.4 Classifier Construction

Goal: to build a classifier able to successfully generalise and distinguish different expressive modes on the basis of their features.

Having successfully acquired a number of features across the high-dimensional space of facial expressions, we then propose to construct a classifier that is capable of both *generalising* the presence of these features across all subjects and instances for a particular expression, and *distinguishing* the variation that can occur within these features that lend them nuance (i.e. can make the same expression look complete different). This stage is therefore very much reliant on the results of the previous stage in selecting such features - and as such, we view the construction of a classifier as an iterative process by which we seek in conjunction for the best features to exploit.

A central concept to appreciate is the existence of the effectual hierarchy within which the perception of human motion can be defined - from low-level movement, to activity, and high-level action [10]. To validate these ideas researchers have looked to applying statistical techniques applied to data from tracked articulated figures (as a form of motion-capture) analysing how people move individually and relative to the context of their surroundings, in order to perform activity recognition [66]. Differing approaches can be adopted, but which still respects the effectual levels of human motion, by instead relying on a more a low-level “blob" based model [83] or by relying on other means of extracting a feature set, such as active appearance model.

Other approaches further this idea by attempting to directly correlate symbolic “verbs”, “adverbs” and “behaviours” with the motion [5, 67, 35, 39]. These adopt a statistical basis, employing a variety of techniques including SVM’s, radial basis function mixture models and HMM’s operating on an annotated training set to establish suitable classifications of motion to be combined. Further unsupervised learning of large motion-captured exemplar datasets can be performed to discover more intrinsic properties of particular movements - to the extent of determining degrees of stylistic elements. Such as the later work by Brand [14], Hack [38] and Bettinger[8] to exploit a HMM based approach on top of their discovered atomic distributions for feature points moving in multi-dimensional “expressive space” - deriving the potential combinations accordingly within the given state at any one instance. This can be done in order to examine how the contributions of various features across a temporal sequence define the classification for the entire “style” or “facial behaviour”.

Our work seeks to build on this, but for more punctuated sequences of key expressions, and with respect to matching classifications to particular Ekman’s Facial Action Units. In doing so, we shall have to localise the classifier to particular regions of the face (depending on the unit), on which it can be trained. This classifier must then be able to learn the variation occurring within these regions that determine the different units locally, and how these units operate together in generating overall expressions. This hierarchical learning structure could be supported by a range of techniques building on the basic concept of a HMM. We intend investigating the various combinations and extensions that can best serve in the construction of such a classifier.

3.5 Generative Synthesis

Goal: to exploit the generative capacity of the classifier in order to create new faces and expressions in 3D.

In having built a suitable classifier, this will in turn support reconstruction of novel data for any desired combination of Action Units, and allow us to alter their variation across key dimensions to define completely new variants and expressions. This can be achieved by utilising the generative properties of "running it in reverse" (taking different probabilistic paths through it) and by altering the distribution parameters to reproduce a feature set of points defining a new temporal sequence. These new points could then be used as *control points* to dictate the movement of an original - or entirely new - 3D face model.

The use of such control points requires us to apply theory from computer graphics, particularly the use of a *subdivision surface* representation of the controlled model (i.e. a neutral 3D face) - progressively defined as a continuous surface by using an increasing level of polygons to represent the detail. This is the most commonly used modern means of supporting complex virtual objects when used in conjunction with *space warping* where the object is not directly manipulated but is effectively controlled by the alteration of the lattice space in which it is embedded - an idea more commonly known as *Free Form Deformation* (FFD). While such a solution can offer smooth global transformations for the whole object, it is often incapable of representing subtle, small scale dynamics - such as facial expression - since the complexity of the control space would then be unworkable or equivalent to the actual object space. Consequently, much research has been focused on finer-grained control and accurate extensions to FFD - which has naturally led to multi-resolution based approaches that allow deformation to be expressed at locally and at different scales [19] (a technique that will hopefully integrate well with our own hierarchy of classifiers for different action units) .

An extension of this (to see the effects of nuance over a greater timeframe) would be to create new "inverse" states that would allow us to perform transitions between other neighbouring states, so enabling creation of longer sequences of expressions produced and cued by the desire to show a particular expressive mode at key points. Considerable research has been carried out in supporting the creation of new and novel movement from exemplar motion capture data sets. This process is generally known as *motion synthesis* and stems from original research in a *warping* approach to smoothly generates transitions in joint parameter curves between sequences [82]. From this has followed a range of other techniques to allow either direct control in *stitching* sequences together by automatically generating further transitions [45], or by using the features extracted for indexing into a database of suitable examples for splicing/blending together [4, 69, 36]. Other solutions generate novel instances in the presence of another control sign - e.g. facial images from novel text sequences (co-articulation) as in [16, 12, 33, 76]. These apply a range of statistical techniques to correlate *phonemes* to *visemes* to form a suitable model that can be applied to generate animation from speech.

3.6 Experimental Validation

Goal: to utilise synthesised models to investigate the subjective and objective validity of nuance and its role in distinguishing realism.

Given that we hope to be able to be able to apply our model of nuance in a generative manner, we shall then be in a position to validate our work. Effectively this is a process by which we wish to establish the merit of any achieved results by *qualitative* (e.g. how good/bad?) and *quantitative* (e.g. how much/less?) means. The goal of which is to answer the research questions phrased in section 1.2, and to ultimately confirm, discard, or reformulate our hypothesis.

From our research question ***"**Is a model incorporating such nuances more realistic than those without?**"*** we propose to define to what extent those elements of our model which constitute nuance contribute to people's perception of expression. This can be achieved experimentally in a similar way to many psychophysical studies, in which we will ask subjects to rate a number of our synthesised models that are created using different parameters. In analysing the results of these *subjective tests*, we can further apply statistical techniques to verify the significance.

Furthermore, it is known in psychology that most expressions do not simply fit to an extreme, but can be a blend of various ones. A certain amount of ambiguity can then arise depending on the range and extent of action units that form the expression. The classic case is between "Surprise" and "Happy" (both of which involve raising the brows, and opening the mouth wide). On this basis, and experiment could proceed to investigate this effect to answer "*What role do particular nuances have in distinguishing expressive modes?*" along the following lines: The model is used to control a "mean" face, in which the classifiers are used to generate a wide range of faces between the two expressions, but by altering the different contributions made by different action units across the whole range of variation. This would produce a number of different sequences which an independent subject can assess for their perceived emotion. In then analysing these results, correlations between particular nuances and expressions would hopefully become apparent.

We also propose more *objective* tests to exactly quantify the operation of the model in realistically predicting the deformation and manner of the human form. This can be done by capturing an additional set of test data (i.e. a different facial expression) which is excluded when constructing the model - that is, *ground truth* or *test set* data. The complete, fused model could then be used to generate an equivalent expression which is directly compared to the test instance to answer "*To what extend can nuances accurately reflect the variations of human expression?*"

The success to which all our experiments contribute will ultimately determine a level by which we can empirically justify the initial claim of our hypothesis. If our model can indeed develop and distinguish key dimensions and variations, which when combined create accurate and realistic results beyond those of previous models - then we will have hopefully shown the importance of the concept of nuance in modelling and accounting for human dynamics.

Bibliography

[1] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[2] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. In *ACM Transactions on Graphics (SIGGRAPH 2002)*, 2002.

[3] M. Argyle. *Bodily Communication*. Methuen and Co., first edition, 1975.

[4] O. Arikan and D.A. Forsyth. Interactive motion generation from examples. *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 21(3), 2002.

[5] O. Arikan, D.A. Forsyth, and J. O'Brien. Motion synthesis from annotations. In *SIGGRAPH '03*, 2003.

[6] N.I. Badler and S.W. Smoliar. Digital representations of human movement. *ACM Computing Surveys (CSUR)*, 11(1):19–38, 1979.

[7] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis, 1999.

[8] F. Bettinger, T.F. Cootes, and C.J. Taylor. Modelling facial behaviours. In *Proceedings of British Machine Vision Conference 2002*, volume 2, pages 797–806, 2002.

[9] R.L. Birdwhistell. *Kinesics and Context*. Penguin Press, first edition, 1971.

[10] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. In *Phil. Trans. Royal Society London*, pages 1257–1265, 1997.

[11] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[12] M. Brand. Voice puppetry. In Alyn Rockwood, editor, *Siggraph 1999, Computer Graphics Proceedings*, pages 21–28, Los Angeles, 1999. Addison Wesley Longman.

[13] M. Brand. Morphable 3D models from video. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2001.

[14] M. Brand and A. Hertzmann. Style machines. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[15] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[16] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proc. ACM SIGGRAPH 97*, 1997.

[17] M. Byun and N. Badler. Facemote: Qualitative parametric modifiers for facial animations. In *IEEE Symposium on Computer Animation*, pages 65–71, 2002.

[18] A.J. Calder, A.M. Burton, P. Miller, A.W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision Research*, (41):1179–1208, 2001.

[19] S. Capell, S. Green, B. Curless, T. Duchamp, and Z. Popović. A multiresolution framework for dynamic deformations. In *ACM SIGGRAPH Symposium on Computer Animation 2002*, 2002.

[20] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. *Computer Graphics*, 28:413–420, 1994.

[21] J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In Eugene Fiume, editor, *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 477–486. ACM Press / ACM SIGGRAPH, 2001.

[22] T. Choudhury and A. Pentland. Motion field histograms for robust modeling of facial expressions. In *Proceedings of the International Conference on Pattern Recognition*, 2000.

[23] E.S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *Pacific Graphics*, 2002.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Proc. of ECCV*, 2:484–498, 1998.

[25] M. Dailey, G. Cottrell, and R. Adolphs. A six-unit network is all you need to discover happiness. In *22nd Annual Conference of the Cognitive Science Society*, 2000.

[26] C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 3rd edition, 1998.

[27] J.W. Davis and H. Gao. An expressive three-mode principal components model of human action style. *Image and Vision Computing*, 21(11):1001–1016, 2003.

[28] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR 2000*, 2000.

[29] P. Ekman. Should we call it expression or communication? In *Innovations in Social Science Research*, volume 10, pages 333–344. 1997.

[30] P. Ekman. Facial expressions. In T. Dalgleish and T. Power, editors, *The Handbook of Cognition and Emotion*, pages 301–320. John Wiley & Sons, 1999.

[31] P. Ekman and W.V. Friesen. *Facial Action Coding System (FACS)*. Palo Alto: Consulting Pshchologists Press, 1978.

[32] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 19, 1997.

[33] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, 38(1):45–57, 2000.

[34] D.A. Forsyth and J. Ponce. *Computer Vision: A modern approach*. Prentice Hall, 2003.

[35] A. Galata, N. Johnson, and D. Hogg. Learning variable length markov models of behaviour. *Computer Vision and Image Understanding: CVIU*, 81(3):398–413, 2001.

[36] M.A. Giese, B. Knappmeyer, and H.H. Bülthoff. Automatic synthesis of sequences of human movements by linear combination of learned example patterns. In H.H. Bülthoff, S.W. Lee, T. Poggio, and C. Wallraven, editors, *Biologically motivated Computer Vision*, pages 538–547. Springer, 2002.

[37] M.A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192, 2003.

[38] C. Hack and C. Taylor. Modelling talking head behaviour. In *In Proc. British Machine Vision Conference BMVC 2003*, volume 1, pages 33–42, 2003.

[39] Y.A. Hicks, P.M. Hall, and A.D. Marshall. A method to add hidden markov models with application to learning articulated motion. In *In Proc. British Machine Vision Conference BMVC 2003*, volume 2, pages 489–498, 2003.

[40] A. Hilton, J. Starck, and G. Collins. From 3D shape capture to animated models. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT 2002)*, 2002.

[41] J. Hoey and J. Little. Representation and recognition of complex human motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 752–759, 2000.

[42] T. Jebara, A. Azarbayejani, and A. Pentland. 3D structure from 2d motion. *IEEE Signal Processing Magazine*, May 1999. Volume 16. Number 3.

[43] O.C. Jenkins and M.J. Matarić. Deriving action and behavior primitives from human motion data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2551–2556, 2002.

[44] D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 2003.

[45] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. In *Proceedings of ACM SIGGRAPH '02*, 2002.

[46] J. Lasseter. Principles of animation. *ACM Computer Graphics*, 21(4), 1987.

[47] J. Lee and S.Y. Shin. Multiresolution motion analysis with applications. In *International workshop on Human Modeling and Animation*, Seoul, June 2000.

[48] C.T. Leonard. *The Neuroscience of Human Movement*. Mosby, 1998.

[49] Y. Li, T. Wang, and H.-Y. Shum. Motion texture: a two-level statistical model for character motion synthesis. In *Proc. of ACM Siggraph '02*, pages 465–472, 2002.

[50] J.J. Lien, T. Kanade, J. Cohn, and C. Li. A multi-method approach for discriminating between similar facial expressions, including expression intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[51] C. Lisetti and D. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the 11 th International Flairs Conference*, 1998.

[52] N. Magnenat-Thalmann, H. Seo, and F. Cordier. Automatic modeling of animatable virtual humans - a survey. In *The 4th International Conference on 3-D Digital Imaging and Modeling*, pages 2–10, 2003.

[53] T. Matsumoto, K. Hachimura, and M. Nakamura. Generating labanotation from motion-captured human body motion data. In *Proc. International Workshop on Recreating the Past - Visualization and Animation of Cultural Heritage*, pages 118–123, 2001.

[54] A. Mehrabian. *Nonverbal communication*. Aldine-Atherton, 1972.

[55] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

[56] T. Nakata. Generation of whole-body expressive movement based on somatical theories. In *Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2002.

[57] J.-C. Nebel. Soft tissue modelling from 3D scanned data. In N. Magnenat-Thalmann and D. Thalmann, editors, *Deformable Avatars*, pages 85–97. Kluwer, 2001.

[58] J.-C. Nebel and A. Sibiryakov. Range flow from stereo-temporal matching: Application to skinning. In *IASTED International Conference on Visualization, Imaging, and Image Processing*, Malaga,Spain, 2002.

[59] B.M. Nigg and W. Herzog. *Biomechanics of the Muscul-skeletal System*. Wiley, second edition, 1998.

[60] M. Pantic, L. Rothkrantz, and H. Koppelaar. Automation of nonverbal communication of facial expressions. In *Proceedings of Euromedia- '98*, 1998.

[61] K. Perlin. An image synthesizer. In *Proc. ACM SIGRAPH '85*, pages 287–297, 1985.

[62] K. Perlin. Real time responsive animation with personality. In *IEEE Transaction on Visualization and Computer Graphics*, 1995.

[63] K. Perlin and A. Goldberg. Improv: A system for scripting interactive actors in virtual worlds. *Computer Graphics*, 30:205–216, 1996.

[64] F.E. Pollick, C. Fidopiastis, and V. Braden. Recognising the style of spatially exaggerated tennis serves. *Perception*, 30(3):323–338, 2001.

[65] K. Pullen and C. Bregler. Motion capture assisted animation: Texturing and synthesis. In *In Proceedings of ACM SIGGRAPH 02*, 2002.

[66] D. Ramanan and D.A. Forsyth. Automatic annotation of everyday movements. Technical Report CSD-03-1262, Division of Computer Science, University of California, Berkeley, 2003.

[67] C. Rose, M.F. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation using radial basis fnctions. *IEEE Computer Graphics and Applications*, 18(5):32–41, 1998.

[68] H. Spies, B. Jähne, and J. L. Barron. Range flow estimation. *Computer Vision Image Understanding (CVIU2002)*, 85(3):209–231, 2002.

[69] L.M. Tanco and A. Hilton. Realistic synthesis of novel human movements from a database of motion capture examples. In *Proceedings of the IEEE Workshop on Human Motion HUMO*, pages 137–142, 2000.

[70] C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding: CVIU*, 80(3):349–363, 2000.

[71] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 6(12):1247–1283, 2000.

[72] D. Terzopoulos and K. Fleischer. Modeling inelastic deformation: Viscoelasticity, plasticity, fracture. *Computer Graphics*, 22(4):269–278, 1988.

[73] F. Thomas and O. Johnstone. *Disney Animation: The Illusion of Life*. Abbeville Press, first edition, 1981.

[74] N.F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):371–387, 2002.

[75] M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96, 1995.

[76] P. Vanroose, G.A. Kalberer, P. Wambacq, and L. Van Gool. From speech to 3D face animation. In *Procs. of the Benelux Symposium on Information Theory*, 2002.

[77] M. Vasilescu and D. Terzopolos. Multilinear analsysis of image ensembles: Tensorfaces. In *Proc. European Conference of Computer Vision*, pages 447–460, 2002.

[78] M.A.O. Vasilescu. Human motion signatures: Analysis, synthesis, recognition. In *Proceedings of International Conference on Pattern Recognition (ICPR 2002)*, 2002.

[79] H. Wang and N. Ahuja. Facial expression decomposition. In *Proc. of the Ninth IEEE International Conference on Computer Vision*, 2003.

[80] K. Waters. A muscle model for animation three-dimensional facial expression. *Computer Graphics*, 21(4):17–24, 1987.

[81] A. Watt. *3D Computer Graphics*. Addison-Wesley, third edition, 2000.

[82] A. Witkin and Z. Popović. Motion warping. *Computer Graphics*, 29:105–108, 1995.

[83] C.R. Wren, B.P. Clarkson, and A.P. Pentland. Understanding purposeful human motion. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[84] Y. Yacoob and L.S. Davis. Computing spatio-temporal representations of human faces. In *CVPR94*, pages 70–75, 1994.

[85] A.W. Young, D. Rowland, A.J. Calder, N.L. Etcoff, A. Seth, and D.I. Perrett. Megamixing facial expressions. *Cognition*, (63):271–313, 1997.