Proceedings: MoMM 2009: The 7th International Conference on Advances in Mobile Computing & Multimedia 14-16 December 2009, Kuala Lumpur, Malaysia

Towards an Interactive Multimedia Experience for Club Music and Dance

Dennis Majoe
ETH Zurich
Clausiusstrasse 59
Zurich
Switzerland
dennis.majoe@inf.ethz.ch

Irena Kulka
ETH Zurich
Clausiusstrasse 59
Zurich
Switzerland
irena.kulka@inf.ethz.ch

Jan Schacher
Zurich University of the Arts
Baslerstrasse 30
Zurich
Switzerland
jan.schacher@zhdk.ch

ABSTRACT

In this paper, we describe completed and ongoing work towards an interactive multimedia system that will appeal to today's youth culture identified as most likely to adopt such novel mobile applications that combines music, dance and technology. We describe our work in the application of three types of successful movement recognition applied in the field of Tai Chi with the objective being to identify gestural primitives of club dance associated with electronic dance music. In this approach, dance movements are first recognized and classified and then mapped, using multiple levels of complexity, to higher level algorithms that can modify multimedia content in real time. The paper describes the mechanisms supporting an attractive alternative to the now standard Video Disc Jockey (VJ) in which members of the dance public are empowered to create multimedia content in real time as opposed to the VJ.

Keywords

Wearable sensors, Gesture Recognition, Dance Analysis, Multilevel Mapping, Real-time Image Control

1. INTRODUCTION

During the 1970's and 80's it was quite common to see attendees at rock concerts holding lit cigarette lighters or candles and waving them so as to participate in the music and social experience. By the 90s and turn of the century the lighters were replaced by bright mobile phone displays lighting up the evening. Similarly the social interaction of the adolescent has grown to include the use of mobile technology as can be seen by the extensive use of messaging services, and now the use of Bluetooth based Social Networking and new Web sites such as Facebook and Twitter, that allow for gossip, friends and new contacts to be made in social environments such as clubs and bars. The use of computer technology in dance and music is now commonplace and in many cases multimedia content is extensively used to support the overall experience at these performances and Video DJ events.

In the home, a major multimedia sector is devoted to entertainment and educational games running on PCs, and consoles such as the SONY PlayStation® and the Microsoft XBOX360®. Recently the Nintendo Wii® made a major impact on interactive games by providing more user interaction with users encouraged to simulate the actions of a sport or dance with even yoga being rolled out as an educative experience. The Wii MotionPlus is bundled as standard with the sports game collection Wii Sports Resort, Tiger Woods PGA Tour, Virtua Tennis, Kidz Sports and similar titles [1]. These applications are being developed at an alarming rate with more and more emphasis being placed on social interaction supported by multimedia interaction.

Meanwhile the use of full body Motion Capture in Virtual worlds and the film industry has generated an expectation in the user community that one day more complex interactions with games and other learning applications will eventually become a reality for the home consumer. The XBOX360 based Project Natal, which has been described as "controller-free gaming and entertainment experience", certainly attempts to remove the tether between the user and the application controller and tries to ensure only body motions are necessary to control the game. Based on infra red cameras and optics the project outcome demonstrates advanced skeletal mapping technology allowing it to track up to four users for real time motion analysis [2].

The use of advanced motion recognition will surely widen as multimedia applications become more mobile and move away from the flat screen interaction paradigm and into the wider environment. Eventually one would expect much higher user autonomy, for example being able to interact with a game while seated with friends in a bar or disco. In this type of mobile context it is more likely the user must wear low cost ergonomic sensors to determine their body motion rather than the external device proposed by Wii and XBOX360.

A similar technological challenge is required to recognize elements of human motion that are far more complex than pretending to play tennis using the Wii controller as a virtual tennis racket. A rich gesture vocabulary may be required to allow human motion to become part of the social interaction and multimedia interaction paradigm. Users are likely to use many sub-movements of dance that are complex but well rehearsed.

Given the sensor and recognition technology, the next requirement would be a mapping system that could assign low level semantics to a wide range of gestural features, and to provide a framework by which these gestural elements could be somehow parsed in order to derive high level commands that kick off functions that generate or modify multimedia content. This mapping would not necessarily be explicit such as one would expect of a mouse click to launch an application. Rather the user may prefer multiple interpretations thereby giving them flexibility in expression. Such socially driven mobile multimedia interaction applications are likely to be adopted and or disseminated by the entertainment community and icons in that community. Today an important source of social entertainment comes from music and dance in clubs and bars in which the Audio DJ and Video VJ take centre place.

2. SENSOR TECHNOLOGY

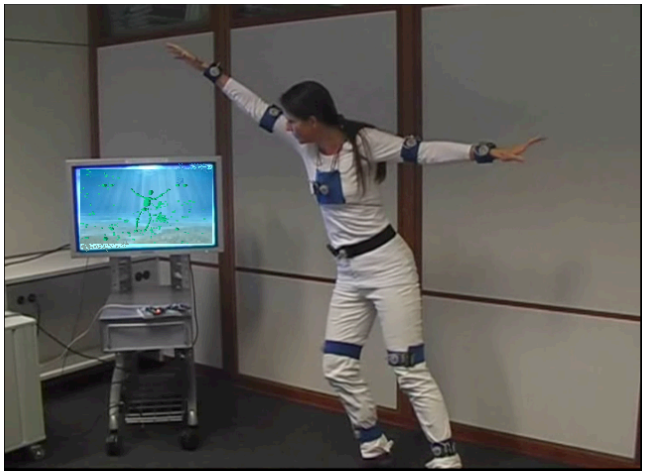

Figure 1: Dancer wearing wireless motion capture sensors

In order to provide human motion sensing on an anytime anywhere basis, there is a need for miniaturized wearable sensors that can sense the movements of the users limbs. Provided these are sufficiently ergonomic (perhaps eventually being part of the clothing fabric) then a user may engage in any normal activities and instantly elect to launch the application. Such a sensor system is being developed as part of this research and although the current status is not ergonomically optimized for futuristic applications, it is acceptable for experimentation with dancers in this multimedia context.

2.1 The Sensors

Our previous work in Tai Chi research [3],[4], was carried out to develop a very low cost, ergonomic motion capture system. The early devices were based on low cost gravity acceleration force sensors however in order to totally measure the student's limb positions a second sensor has recently been added to measure rotations about the vertical axis. In the system developed and reported in [5] the earth's magnetic field provides this second piece of information for sensor fusion.

Further improvements were carried out since then and the prototype sensors used for this reported research have been upgraded so that 10 sensors can be worn on the arms, legs, waist and torso and high speed data delivered to the data processor using a master RF to USB connection. The totally wireless system is designed to be mass produced below a speculated selling price of US$200.00 and to be smaller than a large clothes button. The battery life of each sensor is 24 hours between charges, the sample rate is variable from 10 to 50 Hz and up to 75 users may use the system in the same locality without any conflict. Figure 1 shows a Butoh Dancer wearing the sensors and interacting with a "Swarm Dance" multimedia application. The system is now in use on a daily basis by students in order to develop complex user recognition applications such as Tai Chi and Dance and such a system is considered to be the basis for future commercial systems.

2.2 Related Work

The work reported here is strongly related to previous work conducted in this group in which the movements of a Butoh dancer were studied so as to extract the emotional content of the dance style. In the Butoh Dance project the idea was to recognize or measure the emotional content of the performance by measuring limb acceleration and velocity [6]. Other well known work in this area is attributed to Camurri et al. Camuri has shown that it is possible to establish a correlation between the automatic recognition of emotion expression in dance compared to the spectator [7] and has also developed tools to support the analysis of such emotive expression [8].

However in both of these dance recognition cases the method has relied on measuring flow and energy in movement in order to establish the emotive content. From a sensor perspective this means that the algorithms worked on raw data from the sensors providing indications of acceleration of limbs and rates of turn. In the work presented here however we differentiate from these approaches and prefer to work with the classic idea of gestural primitives of dance. The reason being that in Club Dance we expect that the dancer is automatically constrained to specific modes of expression, specific dance movements, due to the social etiquette that exists in which dance moves are inherited and assimilated into the popular culture. Generally a different sequence of moves is simply repeated without changing it's typical energy and flow profiles. Therefore although we do provide for a low level layer of direct access to raw data and "energy" measurements, the central philosophy is to detect pre-learnt dance moves or poses that are intended to deliver a specific expression.

2.3 Forward Kinematics

As soon as multiple limb sensors are providing information about a user's motion, the potential exists for creating a computer generated avatar that can be rendered to the computer screen and animated in real time by the user. The avatar may be humanoid or of abstract shape, however some model is needed to describe the interconnectedness of the human form using forward kinematics. By generating this type of abstract digital puppetry, multimedia content may be derived with wide entertainment interest.

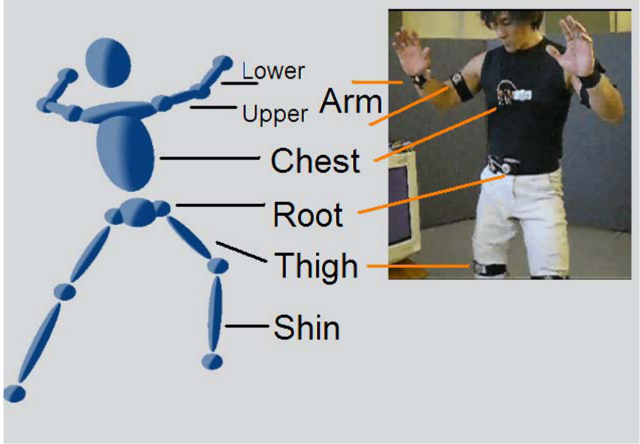

Figure 2: Relationship between sensors and Avatar Bones Model

For the purposes of gesture recognition, when the gesture is simple, such as simple hand movements, it is not specifically necessary to generate an avatar since the raw data from the sensors could also be used directly. However when the whole body is acting as the gesture source, we have found that abstracting the raw data to a higher level, such as the human form, provides much more meaningful data sets to which heuristics can be applied. For example, when one considers that there are at least 10 sensors generating 60 streams of raw data, it is easier to handle a single piece of high level information, for example the 3D Cartesian trajectory of an elbow, rather than each of the 60 signals needed to derive that feature. Figure 2 shows the sensors placed on a user and the corresponding limbs of the digital avatar.

To do this the sensor data is processed and the rotation of the sensor in 3D space is calculated with high accuracy. Since each sensor is attached to a limb, the limb rotation data can be measured and applied within a forward kinematics algorithm utilising an avatar bones model to calculate each limb's Cartesian position in 3D space [9].

The actual absolute position must be estimated by using the rotation data and the length of each limb to calculate each limbs position in 3D space. The position of every part of the body is calculated starting from a root limb such as the waist and hips. The rotation data is applied to this body part to find its orientation in space. Typically the root has two or more connecting limbs. In our bones model the hip is the root, from which the chest and two legs are offspring. The chest connects to the arms and the head. Each virtual limb in the bones model corresponds to a vector displaced from the parent limb from which it has sprung, going all the way back to the root. The rotations are restricted to angles that are typical for a human being. This restricts the avatar's total possible action space. In addition gravity is modelled so that the avatar's feet lock to the simulated ground. Any action such as bending the legs therefore affects the whole body level.

Since we use a standard set of bones lengths in the model, there is no specific relationship between the avatar size and the size of the actual user to whom the sensors are attached. In this way, every user's body can be characterized by this one size fits all model. Therefore the movements of different users can be directly compared despite the fact that actual body dimensions between different users may vary significantly. This technique allows us to calculate the mesh for any number of 3D avatars and render their animation in real time. When the generated avatars are rendered on a computer screen, they may be used in a variety of ways such as in a teaching application which shows digital versions of the users; e.g. Tai Chi teacher and student.

2.4 Tai Chi Basis

Within the range of human motion that may contribute to mobile interactive multimedia it is likely that some pre-learnt movement gestures will be used by people to express themselves. Since Tai Chi includes relatively simple expressive movements that are akin to both dance and martial art, it was chosen as a likely example for gesture recognition.

Tai Chi is made up of a set of specific movements each running in a sequence. For example the most commonly taught simplified 24 movement (Yang) form, may be further deconstructed down to about 70 sequential sub movements each lasting about 2.5 seconds. Each of the 24 movements have names and the work presented here focuses on "Part Mane", "Spread Wings", "Brush Knee", "Repulse Monkey" and "Single Whip". When the sub movement is analysed it may be thought of as a complex human gesture that incorporates the hands, torso and legs moving in harmony.

Therefore as a general model for other applications, Tai Chi provides us with a typical example of basic interaction movements that could be used to drive future mobile multimedia applications.

3. RECOGNITION METHODS

In the work conducted in Tai Chi, three methods have proven very useful under different circumstances. In the first method the static pose of the user is compared with a data base of poses. In the second the dynamic body gesture is classified over a relatively short time interval such as 2 to 3 seconds. In the third the method we classify gestures that have a rhythmic repeatable structure that last perhaps 4 to 10 seconds.

3.1 Static Pose

In the static pose method, the forward kinematics derived avatar provides the 3D Cartesian coordinate points for the generic bones model used to describe all users. Therefore the X,Y,Z positions of the joints or the centre points of limbs may be directly compared between avatars. Two methods were developed on this basis.

For real time applications, the system was developed to accept the input from essentially two users and to generate two avatars side by side. In order to remove any effects of the users facing in different azimuth directions the azimuth of the root hip limbs for both avatars was set equal. Then a simple sum of least square errors was accumulated for the difference in each limb centre point. In addition the error for each limb centre point was tested in order to determine if the one avatar limb centre point was in front, to the side or above the second (reference) avatar's limb centre point. This allowed the system to report the extent of the error in the match of static pose between the two users and also to state which limb was directionally out of phase with the reference avatar. Figure 3 shows the avatars for a teacher of Tai Chi on the left and a student on the right.

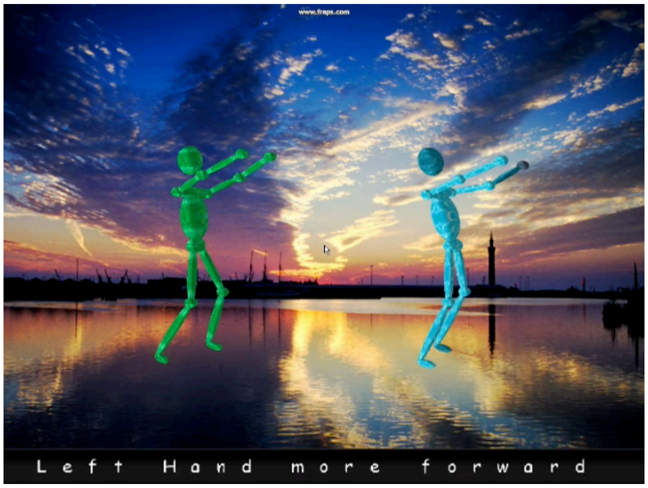

Figure 3: Comparing two Avatars of two users in real time

In order to classify a pose in real time against a data base of poses, the method was extended to incorporate 40 different figurine reference poses. At any time the computer could quickly calculate the difference between the current users pose and the set of reference poses, returning a pose code between 1 and 40, denoting the pose with the least squared error as an indication of best fit.

3.2 HMM for Short Gestures

The Tai Chi movements such as Part Mane and Brush knee take about 10 seconds to complete but they are in fact repetitions of a basic sub-movement. These sub-movements are like the syllables of a longer word. The syllables take short times to complete, last 2 to 3 seconds and cannot be considered as repeating.

For these short gestures the 3D data points from the avatar limbs are used to drive the recognition system comprising a feature extraction layer followed by a Hidden Markov Model (HMM) based classifier.

Several researchers have used HMMs for such gesture recognition tasks. Kahol [10] points out that HMM gesture recognition can use the 3D data in various ways to present gestures as a probabilistic sequence of states associated with the observation of any one of the defined static poses. The HMM method provides two key benefits for gesture recognition of this type. Firstly gestures are not made in exactly the same way each time and within limits the probability of state transitions and emissions within an HMM allow for the generalised characterisation of similar but different motion observation sets even where some small time shifts occur. Secondly the algorithms for the training of the HMM algorithm using training data is now well known [11].

Brand [12] has proposed coupled HMMs that better model multiple interacting processes that have structure in both time and space and has applied it to Tai Chi recognition. In that work the experiments conducted were applied to video based motion analysis and the activity of two arms is studied using the coupled HMM approach. Coupling HMMs requires an insight into the correlations themselves and such a method may represent an impossible task since many different moves have different correlations.

Zhu's [13] ensemble method used 3D features and the key spaces of each human joint to represent human motion. After feature extraction each action class is learned with one HMM and a bagging algorithm, similar to majority voting, is used to ensemble all learners.

After reviewing these alternative approaches it was decided that a mix of the approaches would be implemented rather than any particular one.

3.2.1 Feature Extraction

As with its application in speech recognition or gene sequencing, the HMM approach requires a feature extraction layer that generates abstract codes, an alphabet, that relates to the observation sequence. The least number of codes must express the full range of attributes of the incoming data.

One of the criticisms of using 3D positional data to derive observation codes for human body gestures is that one would need hundreds of observation codes to describe every possible body position. Feature extraction must be performed in such a way that a limited number of different observation codes are needed to represent the overall range of the observation space. Bobick [14] for example used a k-means clustering algorithm to define reference points within the observation space available.

In Tai Chi and sports and dance in general, there is a strong correlation between the positioning of two or more limbs during any sub-movement. Observation coding for multiple limb activity runs the risk of exponential code possibilities unless the dimensionality of the observation space is kept low. Without resorting to complex coupled HMMs, we can combine a hierarchical approach with low dimensionality and linked limb correlations, by using the forward kinematic avatar, to generate one, two or three dimension feature spaces based on combinations of positional information such as hand to hand, foot to foot and hand to foot metrics.

To see how such feature extraction methods could improve recognition a competitive framework was set up in which three different approaches were used to derive feature data, each one becoming more complex in terms of dimensionality and calculation resources. Each feature extraction method would be used to drive identical HMM based machine trained classifiers with the same number of states and the same observation code ranges. Each method would be evaluated based on how often the classifiers provided correct reporting as well as how often false reporting occurred across a number of different Tai Chi movements.

The methods are as follows. The first measures the one dimensional (1D) angle subtended by any two limb ends and the torso, as if the limbs form a scissors. The second method measures the two dimensional (2D) vector made by any two limb ends. Imagine one vector being the line drawn between your two wrists. The third method uses the 3D space positional data of any limb. This is pure 3D trajectory data.

Table 1. Different dimensioned (D=1,2,3) features generated using data from the hands (Lh Rh) and feet (Lf Rf) and torso.

D	A	B	C
1	Angle Lh Rh	Angle Lf Rf	Angle Lh Rf
2	Vector Lh Rh	Vector Lh Rh	Vector Lh Rf
3	x,y,z Lh	x,y,z Rh	x,y,z Rf

Table 1. shows nine feature extraction methods. Within any of the 3 dimensionalities of observation space there are 3 limb combination methods A, B and C that link limb movements directly or indirectly[1]. In all cases two further data manipulations are performed. Firstly in order to ensure that the initial rotational direction of the user is never absolute, at each point in time all of the T samples to be used in a classification are calculated on a differential basis relative to the current first sample. Therefore angular and positional data always appears to run from the same starting point and grows based on the differences from one sample to another.

In practice with the sensor system set to a 10 Hertz sample rate and with each sub movement taking 2.5 seconds to perform, the number of samples taken into the feature extractor, T, is set at 25. For any one capture, the data set is made up of 25 samples of 3D positional data for the legs arms and body.

The HMM state model is not defined manually. After several tests and some tuning a HMM state model with a fixed number of 5 states was chosen. The relationship between the states and their

[1] The forward kinematics always provides a high level indirect limb connection consistency

transitions and the observation emission sequences (allowing for any number of forward and backward state transitions) is left to the HMM Baum Welch algorithm to abstract from the training data. This approach was taken since in our tests we could not choose state models any better than the automatic algorithm. Models with more than 5 states made no improvement in the results.

Training data is produced by capturing a number of repeated versions of the same sub movement in pre code buffers. At this point the 3D data must be translated to observation codes. The number of possible different codes, M, needs to be restricted to a small number, we defined as 12, that adequately describes the regions of the observation space but at the same time is not too large as the HMM algorithm uses up more of the memory and processor time as the values for T, M and the number of Hidden Markov states N increases.

In order to keep M low at 12 the training data generated is subjected to a k-means clustering algorithm to cluster all the training data before we start to generate features. What this implies for method 1, the single dimensional case, is to split the overall range of the angular value to M thresholds along a line. For method 2 it implies splitting the two dimensional space into M areas on the plane and for method 3 it defines M, 3 dimensional zones. Whilst M set at 12 may at first seem a restrictive number of 3D zones, it should be noted that these zones will apply for the data from a specific limb combination and for a specific movement. So even though it is restrictive for other motions of different limbs, it is quite relevant for the specific limb and motion for that particular classifier.

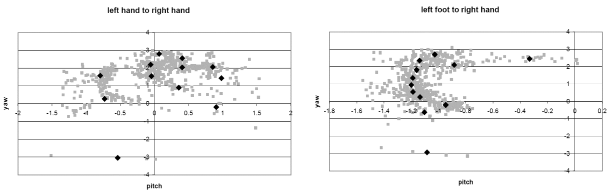

Figure 4: Clustering algorithm derives 12 reference points each defining the centre for an observation code

Figure 4, shows the 12 reference points that are generated from the overall mass of training data. These reference points are used to compare incoming data points. The data point with the smallest linear distance to the reference point is assigned that point's assigned observation code.

The clustering ensures that if on average there are many data points arriving in one region, then more codes are attributed to separate that space. If regions are sparsely populated, then fewer codes are used to specify that space.

3.2.1.1 Data Sets

The HMM approach taken was to train a 5 state HMM with each of the 9 different feature extraction methods. The HMMs would be trained on 5 different Tai Chi movements. This means that we would generate 45 different classifiers with each classifier running with a particular feature extraction method and trained to detect one type of Tai Chi movement.

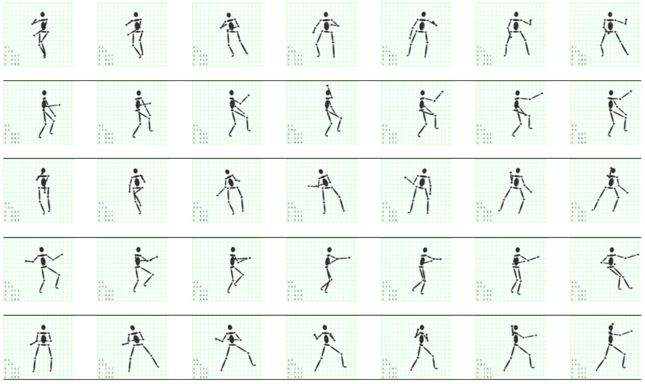

Figure 5: The 5 Tai Chi movements as snap shot Avatars

Figure 5 shows 7 selected shots from each of the 5 moves. Each of the 45 different classifiers was trained and the resulting transition probability matrices were stored. Training data for each movement included 25 repeated movements. Following this a set of 5 movements were also recorded to use as test data. The data was recorded from an average level Tai Chi student.

3.2.1.2 Results

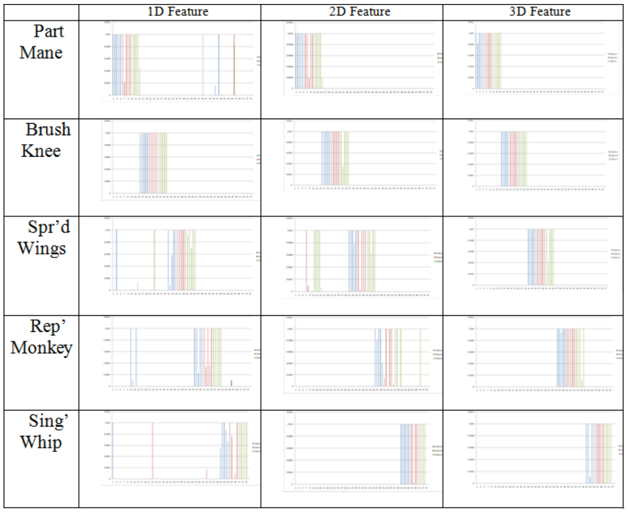

Figure 6: Graphical representation of the Confusion Matrix

Figure 6 shows the forward algorithm classifier outputs for each of the five Tai Chi sub movements in each row. The results are separated into three columns for each of the 1D, 2D and 3D feature extraction methods. It shows which extraction methods perform best and if there is any false positive reporting.

Generally speaking the 3D feature extraction in the right hand column performs the best as it has high forward probabilities for all correctly classified test movements and zero probabilities for all movements not to be classified. All feature extraction methods using x,y,z data are promising. This approach gives a calculated 99.7% recognition rate weighted by probability.

For the middle column, 2D approach, one can see that there is some false reporting particularly for “spread wings” with outputs popping up where not expected; however correct outputs are obtained 86% of the time. The 1D feature approach performs the worst, with false reporting in four of the five movements.

On closer inspection of the results one sees that where one classifier works poorly (e.g. using left hand right hand) the sister classifier (e.g. using the feet) works well. Therefore the 1D and 2D methods can be improved by logically combining related classifier outputs and using majority voting. Then when one classifier fails the partner does not fail, thereby removing the false reports.

3.3 Motion History

Gestures lasting longer than 3 seconds are not appropriate for the HMM approach because for a given sample rate the number of samples used in a classification becomes too high. In the HMM approach, each data sample is encoded into an observation. For any new observation the HMM state machine must compute the probability for remaining or moving to a new state. When observations occur in a sequence that the classifiers state transitions and emissions probability matrices predict well, then the overall probability calculated will be high. However this figure will become smaller as the number of samples increase, despite the fact the observations are matching the classifiers prediction. Hence for very large samples counts, the computed probability for a gesture could be smaller than 10^{-11}. For this reason we restrict the HMM method for use with between 25 and 50 samples at between 10 and 25 Hz sample rates.

Bobick and Davis [11] developed the motion history imaging approach to recognize deeper significance in human motion. In this approach the moving subject is recorded on video or with a time lapse camera so that their motion is seen as a motion blur in the final 2D image. The image obtained has a distinctive shape to it, so for example waving ones arms results in arc shaped sectors. When the black and white image is created in such a way that oldest image contributions are darkest and recent images lightest, then the shaped sectors are observed to have a grey scale gradient from dark to light and this can be used to derive motion when 2D spatial filters are applied.

In our approach we extend this idea by first using the 3D avatar derived from our motion sensors, rather than taking photos of the user, to generate the time lapse history image. Secondly we use the fact that the avatar can be observed from any direction and therefore we can take snap shots of the time lapse rendering from the front view, top view and side view of the avatar. Finally we can conduct many different shading techniques, such that different limbs are presented with different colors and intensities with time.

This 3D approach has several advantages over the 2D approach. The main advantage is that the technique can be applied anytime anywhere using the wearable sensors. There is also no occlusion due to limbs obscuring other limbs.

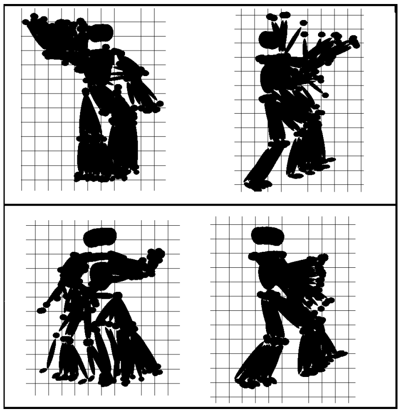

Figure 7: Motion History for Brush knee (Top) and Part Mane (Bottom) movements, front and side views of Avatar

The ability to select camera angle greatly improves the opportunity to improve pattern recognition of particular movements. For example an overhead shot is best for rotations about the vertical axis, impossible with conventional camera methods.

In Figure 7, the top two images are the motion history for the Brush Knee movement, taken from the front (left image) and side (right image). The bottom two images are the history images for the Part Mane movement from front and side. In order to distinguish between these two movements the HMM approach required a significant amount of processing of the sensor data. In contrast when one simply looks at the two history images it is immediately clear that the two movements create different pictures.

Provided that one has enough history and view points, it becomes easier to separate and classify movements using this method. Since the avatars for all users have the same normalised sizes, the history images for different users may be directly compared using simple pixel counting algorithms for pattern template matching and detection.

We therefore found that for long gestures with a highly repetitive content, this approach worked best with simple pattern recognition techniques used to classify the images.

The HMM method succeeds over this approach when gestures are short and there is less history information available with which to build a bold template.

3.4 Recognition Overview

The preceding sections have explained that we have successfully trialed three different modes of human motion recognition. Each mode operates on different time frames of motion data. Each mode of recognition generates a unique classification identifier that demarks a pose, a short gesture or a rhythmic repetitive movement. The techniques work well with Tai Chi and we want to extrapolate this to multimedia interaction in which dancers may interact with multimedia content as a function of their appreciation and interpretation of music.

4. MUSIC DANCE GENRE

There is a very wide choice of music and dance that we could have focused on for this work, however we believe that the most likely early adopters of such technology will be in the young generation of adolescents who enjoy the VJ experience in today's club music dance halls.

Contemporary electronic dance music genres such as Electro House originally started as descendants of disco music in the eighties and developed from genres such as House, Techno, Electro, Trance, Goa Trance, , Progressive House, Deep Trance and Deep House. There appear to be few academic experts in this area and a lot of the research for this reported work arrived from interviews with VJs and dancers. These terms characterize different genres of music, while there are no specific club dance styles. Instead there are numerous variants ranging from repetitive stepping combined with some stereotypical or formally restricted arm gestures to a more inventive gestural vocabulary, sometimes including leg and whole body gestures. For example elements of Hip Hop dance style are sometimes associated with club music and this comprises a lot of stepping, rubber leg bending moves and jumping, with sporadic small gestures in the otherwise rather naturally swinging arms. On the other hand there is the more meditative and inner emotion expressive dance which is associated with Trance and Deep House music. In general the dancing to House and Trance music involves more variation in the arms and freeform kind of movement. Some expressivity can be observed in terms of characteristic spatio-temporal patterns, in terms of the flow of movement (relaxed, vs. flying off, vs. sensual, vs. sharp and aggressive etc), as well as the energetic content (intensity) and direction.

Club music is usually rather simple in structure relying heavily on computer generated sequences of highly repetitive note fragments. There are rarely any significant lyrics and the musical experience is intended to be simplistic yet bursting with emotion and energy or encouraging anticipation and in some cases mystery.

This genre of music is interesting to us because it addresses adolescents keen to adopt mobile technology. It is also interesting because the primitive structural aspects and the raw energy and emotion that is expressed by the music filters into the dance moves and the gestures. That is to say one can observe numerous repetitive rhythmic dance patterns, shorter gestures and some poses that are intended to somehow project expression out to the other dancers. Although there is little expression in terms of articulation of flow and energy, most dancers repeat poses, gestures and rhythmic patterns in combinations that make them individual and self expressive.

Walter Freeman [16] has proposed that dance functions as the "biotechnology of group formation" by offering a means to bridge the gap between self and other. Freeman writes "To dance is to engage in rhythmic movements that invite corresponding movements by others. The reciprocity fosters transcendence over the boundaries of self in physical and emotional communion" Freeman suggested that this primitive attitude to dance existed many thousands of years ago. Since then our brains have not had any reason to evolve away from that trait and even today we are still excited by dancing and by watching dance, particularly of a tribal communal and synchronized nature.

In their review of Motion Perception, Blake and Shiffrar [17] highlight the particularly salient characteristic conveyed by humans being their ability to predict the emotional state of observed individuals through their movements. It is well established, for example, that observers can readily identify the emotion being portrayed by a point light actor given an action as simple as the gesture for knocking on a door. They explain that being able to perceive emotions is part also of being able to express emotions through movement.

We suggest that in a communal situation accompanied by stimulating but essentially primitive music, there are a limited number of 'primitive' psychological motives being predominantly expressed. We can observe phases when the movements of the dancers contains high levels of gestural content whose classification may be achieved using the types of recognition modes described.

Besides of our focus on rhythmic and repetitive gestures, we aim at extracting and classifying spatio-temporal movement patterns that are typical for more abstract, aesthetically and emotionally expressive freestyle forms of dancing.

We believe that various mental movement representations underlie different expressive dance strategies that concretely shape the expression of a motive. This interplay results in the differences in movement that can be characterized by typical sets of movement features.

These signals of communication form the basis for a small vocabulary which may be interpreted by a higher mapping system and which can then launch functions that create or modify multimedia content. The multimedia content will then express the emotions in a manner that feeds back to the dancers as well as provides further entertainment.

The video DJ of today essentially tries to achieve this by his own interpretation of the music and by modifying multimedia content explicitly. This means that any VJ presentation is purely subjective and individual. What may be required for a wider commercial service is an empowering mechanism that places control in the hands and legs of the public dancers.

5. MAPPING

The ongoing multimedia interaction research we are conducting takes the promising results of the work with Tai Chi recognition and applies it to interactive multimedia. The intention is to trial a system under development whose aim is to essentially automate the Video DJ's work processes and at the same time empower the dancing audience by allowing them to mediate the multimedia experience as a function of their dance expression.

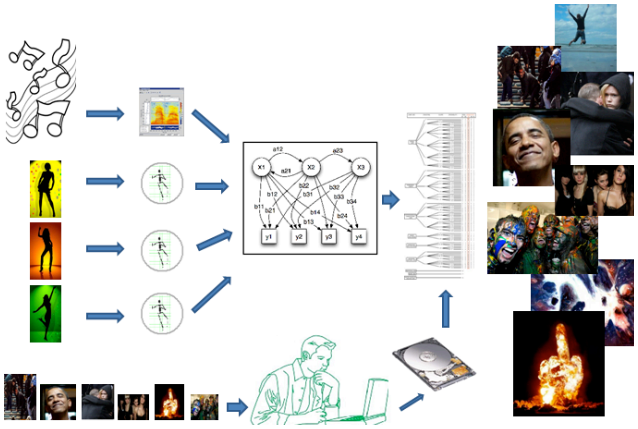

Figure 8: Overview of the VJ system with music, dancers and multimedia on the left and final content output on the right

The system being constructed and depicted in Figure 8 will receive dance motion information from three dancers wearing the motion capture sensors. In addition the music being played will be processed to extract the fundamental beat of the music.

The recognition system will combine the three recognition modalities using the music beat as a sample timing reference. In fact the motion capture sample rate is made adaptive and set at several times the detected beat rate. Clearly there is a great deal of extra information in the music that one would like to classify in order to provide further dance interpretation. Music categorization has become a formidable research area given the amount of music now available on the Internet. Silla et al. [18] for example have been able to process music to classify Latin Dance music with high accuracy. Although being able to automatically interpret club music may add to our multimedia interaction system, currently such methods fall outside our current scope but will be considered in future.

Off line the VJ will compile a large amount of multimedia content, primarily still images chosen for their powerful expression. The VJ will add metadata to each of the content fragments, where this metadata comprises relevant keyword descriptors from a pre-defined list.

A mapping layer will take as input the recognition results from the pose, short gesture and rhythmic movement recognition layer and estimate what multimedia content is relevant and what functions should be applied to the content in preparation for and during its display. Functions include for example the timing, sizing, positioning, movement and color rendering of an image.

5.1 Realtime Images

In the last decade real time composition of images, be it photographic, cinematic or graphical has become a standard practice. Digital processes have enabled the use of instrument-like interfaces or algorithmic, rule-based systems on a level of detail and complexity sufficiently rich for artists to explore. The language of abstract moving images and generative art goes back almost to the beginnings of moving images themselves. Pioneers of the field of expanded cinema [19] used traditional technological crafts to produce their work. Video-art and live cinema further developed the methods and styles in the direction of visual music [20],[21]. In parallel the popular culture of video-jockeying evolved in clubs, where the context and intention of the work has a radically different focus. The main difference between the two fields might be seen on a semantic level; the language and the elements used follow divergent trajectories, on the one hand exploration of immersive audio-visual spaces and on the other hand the enhancement of the club experience by expressive image-sound relationships.

In practical terms a distinction between two types of methods of producing real time images can be made. On the one hand the live-performance instrument is derived from a practice of live-electronics in music. Gestural and parametric controllers such as keyboards, mixing desks etc. are used by the artist to shape and control the images in real-time. On the other hand the algorithmic system generates structure and flow from a set of rules and an input of meaningful data of some kind (usually motion capture or gestural interaction, but also data of any kind, i.e. scientific data visualization). In the case of interactive installations [22] audience action is translated to visual (and aural) output via an interpretation algorithm, which formalizes relationships between the sensed action and the language of the resulting media (sound and image). Since in the physical world there are usually not many direct correlations between a mode of interaction and the media process used to express it, the translation process involves the creation of a symbolic abstract aesthetic model that integrates overarching artistic principles.

5.2 Mapping and the Underlying Principles

In order to produce a richly textured and interesting visual output that still exposes the initial gesture impulses driving it, a mapping strategy has to be devised which allows for hierarchical layers of expressive information to be processed. [23] Parallel processing of these streams in a variety of ways and in increasing complexity enables a multifaceted expressive result.

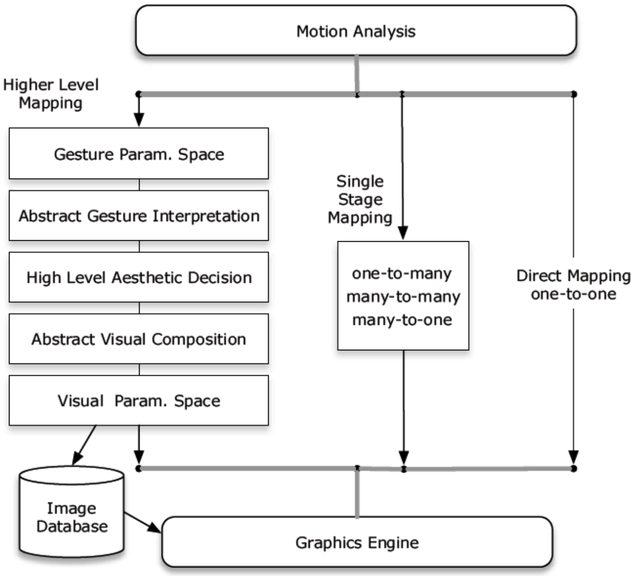

Figure 9: Three levels of Mapping

Figure 9 shows three levels of mapping; direct mapping, single stage mapping and high level mapping.

On the lowest level direct physical parameters obtained from the motion analysis, such as acceleration, speed, absolute or relative position can be assigned to expressive elements in the image control. This generates a sense of synchrony, which is important in strongly rhythmical dance.

An intermediate level is structured in such a way as to combine a number of parameters into new streams of information, which express cognitively more complex but essential relationships [25]. These streams can be applied to higher-level descriptors of the image processes, which in turn get translated to several types of visual elements simultaneously. For example several specific linked gestural types may be known to express sensuality or aggression. In this case it seems prudent to use an explicit set of functions that will collate relevant content accordingly.

The highest level of mapping attempts to move into an abstract domain detached from individual parametric entities. Through the use of several layers of abstraction (cf. schema) salient features of the motion domain are reinterpreted as semantic units, such as gestures. These elements are combined in an algorithm, which operates on aesthetic categories/dimensions and produces conceptual compositional decisions and control streams rather than direct mapping connections. These high level descriptors are subsequently translated into abstract graphical controls, which in turn generate parametric controls and discrete events necessary to produce desired images. These multiple layers are depicted in Figure 10.

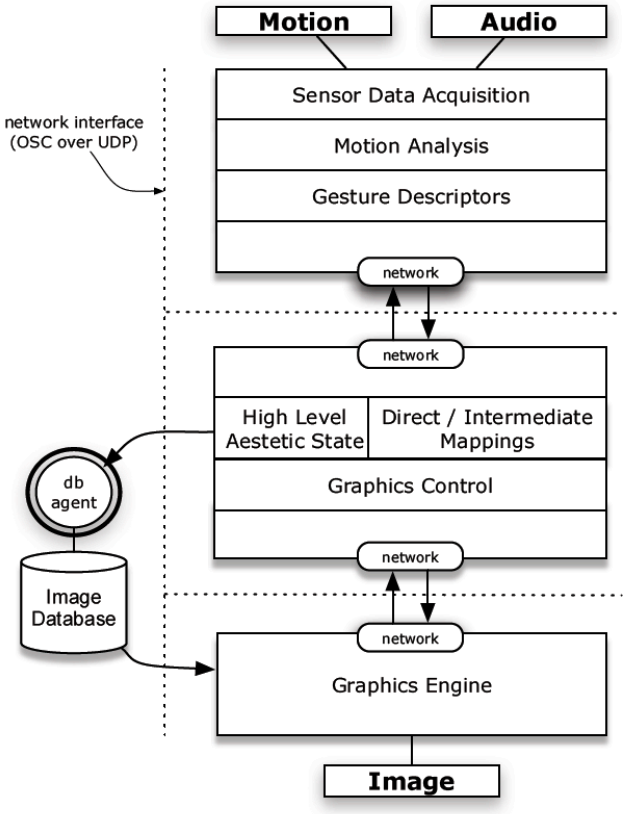

Figure 10: Multiple layers that allow complex interpretation

The combination of several layers of complexity allows for an aesthetically satisfying visual output while maintaining a recognizable link to the primary impulses by the dancers. It must be stressed that the lack of a clearly defined and formally recognized correlation between dancing and image-performance implies a certain amount of subjective and artistic interpretation. This might seem problematic but can also be considered beneficial to the project. Finding a balance between structurally strongly linked motion elements and aesthetically motivated decisions is crucial. This is mitigated by the fact that the two visible elements of the Dance VJ situation, the dancers and the projected images will be perceived simultaneously (by the dancers but also by outside spectators) and will form an expressive whole, comprised of imagination both in motion and in image, part of it linked through the system, part of it juxtaposed in people's perception.

6. CONCLUSION

The exploitation of mobile interactive multimedia in the near future requires the identification of a relevant segment in the community who may be open to the concept and a path to exploitation that is based on substitution from current modalities. In this paper we argue that the use of multimedia content in dance has a powerful attraction to today's youth culture and that the Video DJ role is a good starting point for combining the role of the dancer and the active creation of multimedia content.

We have reported on results obtained with three types of human motion recognition that are ideal for the recognition of dance, the choreography of which includes significant basic or primitive self expression as well as communal expression.

We have then described our work towards an integration of motion recognition and mapping which follows a specific architecture that allows for expressing multiple levels of complexity.

7. ACKNOWLEDGMENTS

Our thanks go to Professor Juerg Gutknecht for his comments and suggestions in framing this research and software support from Lars Widmer and Philip Tschiemer.

8. REFERENCES

[1] http://en.wikipedia.org/wiki/Wii_MotionPlus/.

[2] http://en.wikipedia.org/wiki/Project_Natal/.

[3] K. Kunze, M. Barry, E. Heinz, P. Lukowicz, D. Majoe , J. Gutknecht, "Towards Recognizing Tai Chi - An Initial Experiment UsingWearable Sensors", 3rd International Forum on Applied Wearable Computing, Mobile Research Center, Universität Bremen, Bremen, Germany, 2006.

[4] D. Majoe, I. Kulka, and J. Gutknecht, "Qi energy flow visualisation using wearable computing," Pervasive Computing and Applications, 2007. ICPCA 2007. 2nd International Conference on, pp. 285–290, 26-27 July 2007.

[5] D. Majoe, M. Estermann, N. Ranieri, J. Gutknecht, "Ergonomic Low Cost Motion Capture for every day health exercise" Pervasive Computing and Applications, 2008. ICPCA 2008. Third International Conference Volume: 2, On page(s): 627-6326-8 Oct. 2008.

[6] Jürg Gutknecht , Irena Kulka, Paul Lukowicz, Sven Stauber, Tom Stricker, "Advances in Expressive Animation in the Interactive Performance of a Butoh Dance ",

Communications in Computer and Information Science, Transdisciplinary Digital Art. Sound, Vision and the New Screen Volume Volume 7 , SSN 1865-0929, Springer Berlin Heidelberg, 2008.

[7] Antonio Camurri, Ingrid Lagerlof, Gualtiero Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques.", International Journal of Human-Computer Studies In Applications of Affective Computing in Human-Computer Interaction, Vol. 59, No. 1-2. (July 2003), pp. 213-225.

[8] Antonio Camurri, Barbara Mazzarino, Gualtiero Volpe, "A tool for analysis of expressive gestures: The EyesWeb Expressive Gesture Processing Library.", InfoMus Lab (Laboratorio di Informatica Musicale) DIST – University of Genova, http://infomus.dist.unige.it

[9] Wikipedia, "Forward kinematic animation", http://en.wikipedia.org/wiki/Forward_ kinematic_animation

[10] Kanav Kahol, Priyamvada Tripathi, Sethuraman Panchanathan, " Computational Analysis of Mannerism Gestures." Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Pages: 946 – 949, 2004.

[11] L.R. Rabiner. "A tutorial on HMM and selected applications in speech recognition". In Proc. IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[12] M. Brand, N. Oliver, A. Pentland, "Coupled hidden Markov models for complex action recognition." Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997.

[13] HongLi Zhu, Peng Ying Du, Jian Xiang. "3D Motion Recognition based on Ensemble Learning", Proceedings of the Eighth International Workshop on Image Analysis for Multimedia Interactive Services, 2007.

[14] AF Bobick , AD. Wilson, "A state based technique for the summarization and recognition of gesture.", Proceedings 5th international conference on Computer Vision 1995.

[15] A. Bobick and J. Davis. The recognition of human movement using temporal templates. IEEE Trans. Patt. Analy. and Mach. Intell., 23(3), 2001.

[16] Freeman, W.J.," Societies of Brains. A Study in the Neuroscience of Love and Hate" Hillsdale, NJ: Lawrence Erlbaum Associates 1995

[17] Randolph Blake, Maggie Shiffrar, "Perception of Human Motion ", Annual Review of Psychology, Vol. 58: 47-73 (January 2007)

[18] Silla, C.N.; Kaestner, C.A.A.; Koerich, A.L."Automatic music genre classification using ensemble of classifiers", IEEE International Conference on Systems, Man and Cybernetics, 7-10 Oct. 2007 Page(s):1687 – 1692

[19] Youngblood, G "Expanded Cinema", P. Dutton & Co., Inc., New York 1970

[20] M. Makela, "Live Cinema: Language and Elements", MA in New Media, Media Lab, Helsinki University of Art and Design, 2006

[21] J.C. Schacher, "Live Audiovisual Performance as a Cinematic Practice", The Cinematic Experience, Sonic Acts XII, February 2008 Amsterdam, Sonic Acts Press

[22] J.C. Schacher, "Action and Perception in Interactive Sound Installations: An Ecological Approach", Proceedings, New Interfaces for Musical Expression, NIME 2009 Conference, Pittsburgh, PA

[23] A. Camurri, "Multimodal Interfaces For Expressive Sound Control", Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04), Naples, Italy, 2004

[24] Camurri, A, Mazzarino, B, Volpe G, "Expressive Gestural Control Of Sound And Visual Output In Multimodal Interactive Systems" Proceedings of the Sound and Music Computing Conference SMC 2004, Paris, France

[25] J.C. Schacher, "Gesture Control of Sounds in 3D Space", Proceedings, New Interfaces for Musical Expression, NIME 2007 Conference, New York, NY