ATRテクニカルレポート表紙

[公開]

TR-M-0043

Video Rhythm and Motion Analysis

マ-カス　ザ-キイ
Marcus CSAKY

鈴木　良太郎
Ryotaro SUZUKI

朴　鍾一
Jong-Il PARK

井上　正之
Masayuki IINOUE

1999.3.24

ATR 知能映像通信研究所

Video Rhythm and Motion Analysis

Marcus Csaky, Ryotaro Suzuki, Park Jong-Il, Masayuki Inoue
ATR Media Integration & Communications Research Laboratories

Video Rhythm and Motion Analysis

Marcus Csaky, Ryotaro Suzuki, Masayuki Inoue
Media Integration
Communication Lab, ATR
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
Email: {mcsaky,ryotaro,inoue}@mic.atr.co.jp

Abstract

Three different analysis techniques for video rhythm and motion analysis are presented. First, the spatial frequency analysis of movies is considered. Two sequences from different movies are compared subject to a variety of criteria with this method. A difference between the two sequences is characterized and various features are identified within each media. These features include scene changes, close-up movie shots, and long movie shots. Second, the principle component analysis of dance optical flow is considered. The optical flow of a segmented dance video is considered. A dance composition is segmented into parts that differ in terms of expression as displayed by the dancer. Each segment's respective motion data is reduced by principal component analysis (PCA). These principle components are compared to the principal components generated from a subjective experiment in which the subjects commented on parts of the dance composition. Lastly, a dance rhythm computation system is considered. A real-time system is developed in order to estimate a user's dance rhythm. The user views a skilled dancer and then attempts to duplicate this "training" dance rhythm. The system provides feedback to the user to indicate the closeness of rhythm attained. The rhythm calculation consists of a maximum entropy method (MEM) windowed frequency analysis of binary image differenced (BIN) data. Multi-user capability of the system is also explored.

1 Spatial Frequency Analysis Preliminaries

The method proposed here is in extension to the "Image Wave" project carried out in the Media Integration and Communication Lab at ATR. This project pursues the physical attainment of rhythm information from movie clips in order to realize the automatic synchronization of multimedia content creations[1]. With this method it is hoped that one can characterize wave-like structures within particular media contents (the media content here will always refer to a whole or part movie). Furthermore, it is proposed that these wave-like structures in movies will yield specific movie wavelengths (or frequencies). One idea being investigated in the Image Wave study is the combining of more than one movie. A simple example is to have one movie running in the foreground and a different movie simultaneously running in the background. After obtaining the aforementioned wavelength it would be easier to coherently blend multiple movies together due to the fact that movies with similar rhythm could be classified and synthesized.

Attaining the desired movie wavelength is the end result in this study. At first it is necessary to devise a method of sampling movie information and of getting some indication of what type of movie the particular selection is. The method of spatial frequency analysis is selected due to its frequency domain representation. This method is by no means the only one available and in fact similar methods may be easier to use and to interpret the results of. For instance in the results section, some spatial frequency derived results are compared to results obtained by two other methods (the change in the average value of the brightness and the change of the variance value of the brightness).

The two movie segments analyzed (expediently named Odessa1 and Odessa2) are both scenes from the famous Russian film "Battleship Potenkin" by S.M. Eisenstein. These two segments are selected because of their vast differences. Odessa1 is "very slow" moving with a lot of long shot camera styles whereas Odessa2 is very quickly paced and it has a lot of close-ups and scene changes. Figure 1 are typical movie sequences from Odessa1.

Each progressive image in these 2 image sequences is sampled at a one second interval. Note that there is very limited movement during each 3 second section of footage in the left and right sequences. Figure 2 are typical movie sequences from Odessa2.

These 2 image sequences are again sampled at one second intervals. Note that there is a considerable amount more movement during the 3 second sequences of footage in figure 2 than there is in the sequences in

Fig. 1: Odessa1 sample scenes

Fig. 2: Odessa2 sample scenes

figure 1. Watching the movie segments makes it clear that these two selections create a different mood in the viewer's mind. The goal of this analysis is to be able to differentiate these two selections quantitatively.

1.1 Method

The temporal change of images in movies can be analyzed in a variety of different ways. Some of these methods include the change of the average value of the brightness, the change of the variance value of the brightness, wavelet analysis, and spatial frequency analysis[1]. The change of the average value of the brightness can be calculated as follows:

$$\overline{P}_{i+1} - \overline{P}_i \tag{1}$$

where $\overline{P}_{i+1}$ is the average value of the pixel brightness for the $_{i+1}$st image (or frame). The change of the variance value of the brightness can be calculated as follows:

$$\frac{\sum_i (P_i - \overline{P}_i)^2}{n} \tag{2}$$

where $\sum_i$ is the summation over all the image pixels for the $_i$th frame, P_i is the pixel intensity for the ith frame, and $\overline{P}_i$ is the average pixel intensity for $_i$th frame.

Spatial frequency can be directly obtained from the two-dimensional (2D) Fourier transform. An image can be represented by a 2D function $f(x, y)$, where x and y are the horizontal and vertical orientations respectively within the image, which gives an indication of the image brightness at the point (x, y). In this study, the magnitude of $f(x, y)$ is the Y (brightness) value from the YUV gray scale scheme. The spatial domain representation, $f(x, y)$, can be transformed into the spectral domain $F(u, v)$ via the 2D Fourier transform. In order to represent this spectral domain (u, v) in one dimension, the normalized spatial frequency is defined as:

$$\sqrt{\left(\frac{u}{X/2}\right)^2 + \left(\frac{v}{Y/2}\right)^2} \tag{3}$$

where $X/2$ and $Y/2$ are the magnitudes of half of the horizontal and vertical image boundaries respectively. Due to the periodicity properties of the Fourier transform[4]:

$$F(u, -v) = F(u, Y - v) \qquad F(-u, v) = F(X - u, v)$$
$$F(-u, -v) = F(X - u, Y - v)$$

it is only necessary to consider one quarter of the u and v values. There is much literature available concerning Fourier transforms and spatial frequency[2, 3] but basically spatial frequency gives an indication of the nature of change throughout an image or series of images. Higher spatial frequencies represent rapid intensity changes in an image whereas lower spatial frequencies represent smoother image intensity changes [2].

C software was written to decompose a movie selection into its individual frames. Each frame's brightness values were saved as a 2D real-valued matrix. Using the Matlab function call "fft2", a 2D Fourier transform was taken of the data to yield a complex-valued 2D matrix of data. From this 2D matrix ($M_{i,j}$) the power is calculated to be:

$$P_{i,j} = |M_{i,j}|^2 \tag{4}$$

were $|M_{i,j}|$ is the absolute value of the $M_{i,j}$th component of the matrix. This data was then raster scanned to associated each power value with its corresponding spatial frequency.

Some tests were conducted in order to better interpret the results of this spatial frequency image analysis method.

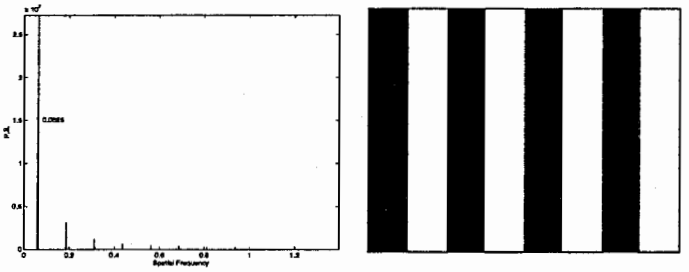

Fig. 3: Spatial frequency of vertical line pattern

The plot on the left of figure 3 is the spatial frequency analysis of the simple vertical line pattern on the right of figure 3. The dimensions of the vertical line image is 128x128 and the width of each vertical bar is 16 pixels. The first maximum power spectra value in the left plot occurs at a spatial frequency of 0.0625. The corresponding period (period = $\frac{1}{frequency}$) is 16 which corresponds to the period of the repeating black and white barred pattern. The small peaks between the two power spectra maxima are due to the lack of smooth transition between the black and white vertical lines. Figure 4 was considered next in order to improve upon the colour discontinuities between the black and white bars.

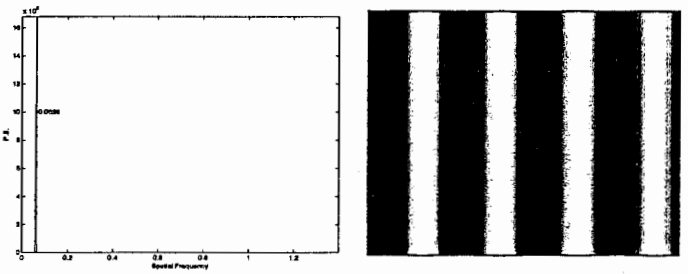

Fig. 4: Spatial frequency of vertical shaded line pattern

The image on the right of figure 4 was created by plotting the brightness intensity of a sinusoidal wave pattern and the plot on the left of figure 4 is the spatial frequency analysis of this image. One can again see the expected maximum at the specific spatial frequency (0.0625) which corresponds to a horizontal period of 16. However, the subsequent power spectral lines are eliminated due to the smoothness of the transition from black to white vertical lines. The left spatial frequency plot was made of the circular pattern in figure 5 in order to test the circular symmetric nature of the overall spatial frequency.

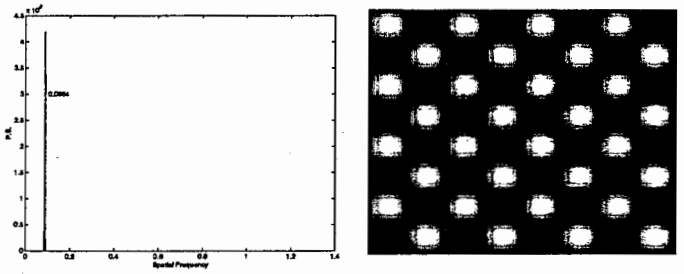

Fig. 5: Spatial frequency of circular symmetric pattern

The image on the right of figure 5 was made by plotting the brightness intensity of the product of a horizontal sine wave and an inverted ((-1)*sinewave) vertical sine wave of a image coordinate. The power spectra maxima at the spatial frequency of 0.0883 corresponds to $\sqrt{0.0625^2 + 0.0625^2}$. With this methodology in place, it is possible to analyze the spatial frequency data from the two movie segments (Odessa1 and Odessa2).

1.2 Results

The overall goal in this spatial frequency analysis is to characterize a rhythm (wavelength or frequency) in a movie. To attain this goal, the movie clips (Odessa1 and Odessa2) are qualitatively analyzed by three contrasting methods. The first method considered is comparing the two time series of the average value of image brightness between frames of Odessa1 and Odessa2.

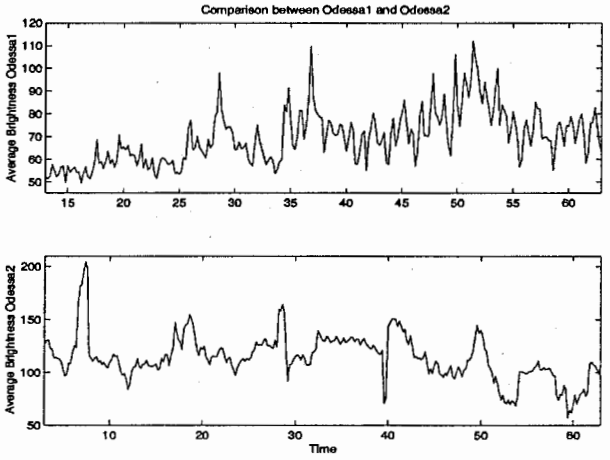

Fig. 6: Average frame brightness change

The two plots in figure 6 represent the average brightness change from frame to frame in Odessa1 and Odessa2. The x-axis is the temporal progression of the movie segment and the y-axis is the average brightness of the frame at a particular time. These plots represent about a minute's worth of data and it is clear that the structure of the two graphs is very different. The bottom graph is smoother and marked by distinct events. An event can be any noticeable feature within the media such as a scene change, a close-up, or a longer camera shot. Event transitions are easy to spot in the second plot as they occur at places of rapid changes in average frame brightness (approximately 7, 28, and 39 seconds for example). Although this qualitative analysis is crude, it is easy to see the difference between Odessa1 and Odessa2.

The second method considered was comparing the two time series of the variance value of the image brightness from frame to frame in Odessa1 and Odessa2.

In figure 7, the two plots represent the change of the variance value of the brightness from frame to frame in Odessa1 and Odessa2. The x-axis is the temporal progression of the movie and the y-axis is the variance value of brightness of the frame at a particular time. Again, these plots represent one minute's worth of data and the structure of the two graphs is very different. The top graph has more noise associated with it than does the bottom graph. Events can be clearly located in the bottom graph (for example, from about 19 to

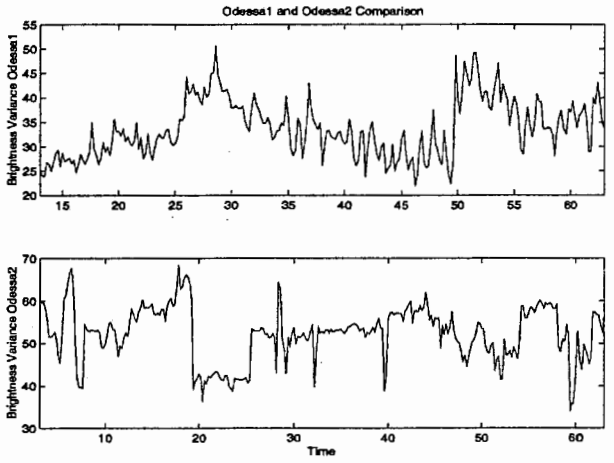

Fig. 7: Variance frame brightness change

25 seconds and from 32 to 39 seconds). The bottom plot suggests some rhythm from about 20 to 60 seconds because after approximately each 6 second segment there is a change in the time series. The same type of analysis was carried out utilizing the spatial frequency analysis method.

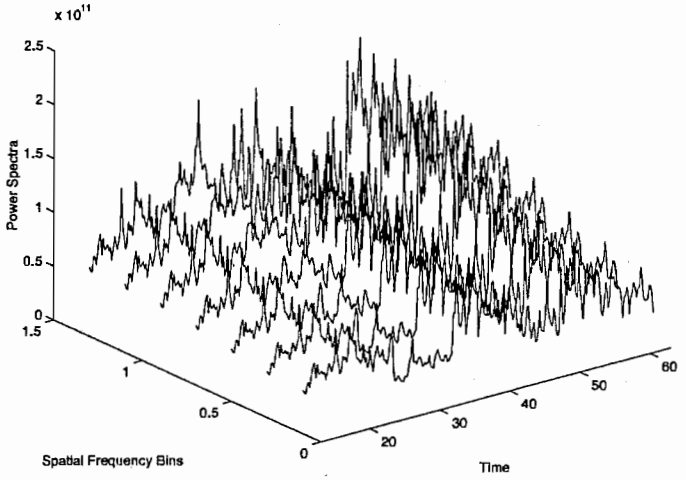

Fig. 8: Spatial frequency for Odessa1

Plot 8 represents data from Odessa1 (a similar plot is made for Odessa2 although not included here for space considerations). The x-axis is the temporal progression of the movies. The y-axis is binned spatial frequency (in bins of width 0.2). An example of this binning method is that one time along the 0.1 spatial frequency path represents the sum of all the frequency components between 0.0 and 0.2 at this time. This binning method was necessary in order to coherently display the large amount data. The z-axis is the power spectra. With this method it is not easy to visualize the data nor is it easy to see any differences between Odessa1 and Odessa2.

There seems to be a certain contradiction with the spatial frequency results obtained and those that are expected in a natural setting. Generally low frequency components are usually more prevalent in nature than high frequency components. However, the human visual processing system is complex and perhaps not yet fully understood. An interesting topic is the question of what form image information takes in the visual system. One theory is that image information in the brain is simultaneously stored in various parts of the brain in both the time and the spectral domain[5]. It is presently not possible to exactly duplicate this process with simple computer based vision models such as the one presented here and hence this may partially explain the difference between some natural processes and the same process viewed through computer vision techniques.

It was decided to select the two lowest and highest frequency bins in order to analyze the movie trends. In theory, the high and low frequency bands should display the biggest difference between the two movie segments due to the fact that the low frequency changes are representative of smooth image changes while the high frequency changes indicate rapid intensity changes. Therefore, more specific data was plotted in order to further analyze figures 7 and 8.

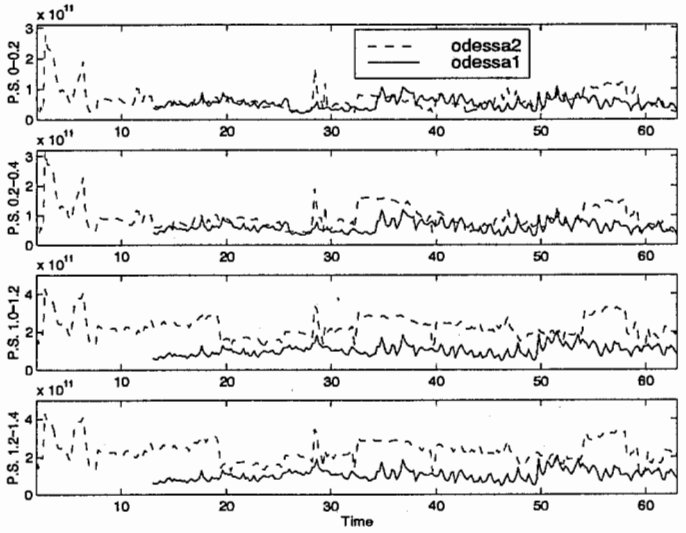

Fig. 9: High and low frequency movie comparison

In the plots of figure 9, the data corresponding to Odessa2 begins at about the 2 second mark while the Odessa1 data begins at about the 13 second mark (the Odessa1 movie clip does not begin until this time). The top two plots represent the two lowest frequency bands while the bottom two plots represent the two highest frequency bands. The x-axis is the temporal progression of the movie segment and the y-axis is the power spectra at each frequency bin. At the lowest frequency bin, the magnitude of Odessa1 and the magnitude of Odessa2 are about the same. However, at the two highest frequency ranges it is clear that the magnitude of the power spectra of Odessa2 is much larger than that of Odessa1. Hence, both movie segments have approximately the same amount of lower frequency components while Odessa2 has a greater higher frequency component. Theoretically, one would expect the low frequency component of Odessa1 to greater than the low frequency component of Odessa2.

Therefore, it is possible to characterize the difference between Odessa1 and Odessa2 by using the spatial frequency method. The next question was why does this method work? It is obvious that by using spatial frequency, one can detect scene changes within a movie but can one tell the difference between camera angles and close-ups versus long shots? Figures 10 and 11 are two examples of close-ups versus long shots. These two sets of images were compared using the spatial frequency method.

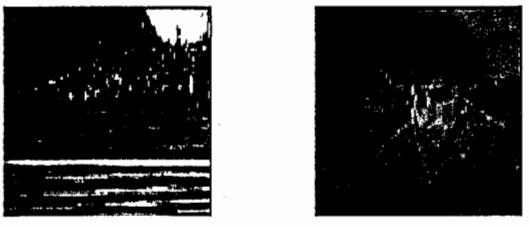

Fig. 10: Odessa2 long shot and close-up example 1

Figures 12 and 13 represent the spatial frequency anaylsis of the images in figures 10 and 11. The top two plots in figures 12 and 13 display the spatial frequency along the x-axis and the power spectra for the particular

Fig. 11: Odessa2 long shot and close-up example 2

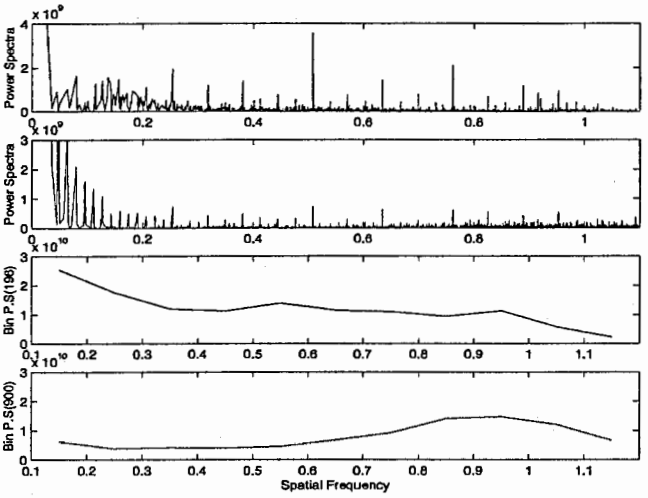

Fig. 12: Spatial frequency comparison of long shot and close-up example 1

close-up or long shot image along the y-axis. The bottom two plots in figures 12 and 13 have bins (where the frequency binning was done similarly to the frequency binning of figures 7 and 8) of spatial frequency along the x-axis and the power spectra along the y-axis. In figure 12, the third plot from the top is the binned spatial frequency representation of a close-up shot (right image in figure 10). One can see that the long shot image has a greater magnitude of power spectra at lower frequencies than does the close-up image. Furthermore, the close-up image has a greater magnitude of power spectra at higher frequencies than does the long shot image. This corresponds to the fact that higher spatial frequencies represent rapid intensity changes in an image whereas lower spatial frequencies represent smoother image intensity changes. A similar trand can be noticed in the analysis (figure 13) of figure 11. It is worthy to note that these results are obtained with specifically selected image combinations and perhaps the outcome would differ if these ideal conditions where not used.

2 Physical and Psychological Analysis of Dance Preliminaries

The purpose of the open house dance study is to derive a measure of the correspondence between the observed psychological and the measured physical aspects of a dance performance. This report mainly deals with the physical data analysis of the dance motion. The motion of a pre-recorded dance performance is analyzed. This particular dance segment was selected due to it variety of movements, tempos, and emotional impact. The dance is performed (Aimi Hara, Kobe University) and directed (Mariko Shiba, Kobe University) so that the range of movements varied from slow and methodical to explosive. The dance sequence is classified into seven segments (Junko Tsukamoto, Tenri University), or motifs, where each motif demonstrates a different emotional expression. These seven emotional segments are subjectively chosen but the goal is to characterize the similarities between this subjective analysis, or the psychological (as described later), and a physical analysis. The physical analysis consists of the decomposition of the dance movie into individual image frames and then the subsequent optical flow analysis of successive dance image frames. The method of physical motion analysis used in this study is very similar to the method used in[6] to analyze simple walking motion and the lip motion that occurs during speech. Although the algorithm is the same, the motion ranges that are encountered

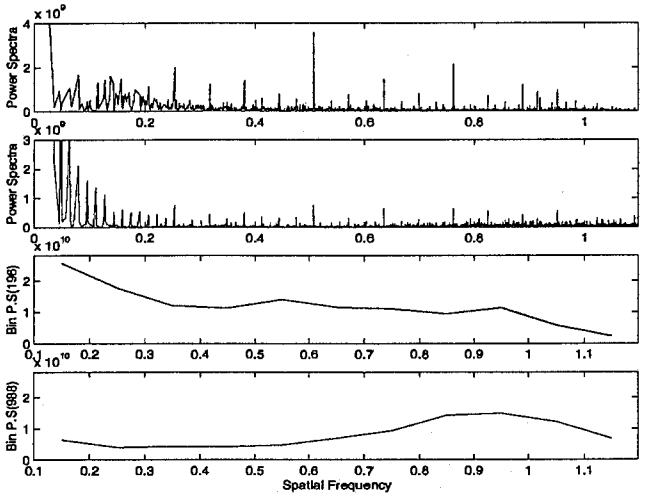

Fig. 13: Spatial frequency comparison of long shot and close-up example 2

in the dance performance are different, and perhaps more complex, that those of Black's analysis.

The number of data points obtained from the optical flow analysis is equal to twice the resolution of the image (or $2 \times n \times m$ where n and m are the width and the height of the image respectively). The speed of a normal movie is 30 frames per second. The video sequence considered is approximately 80 seconds in length and therefore, it contains a very large set of data to analyze. The method of principal component analysis (PCA) is utilized to more easily visualize this large data set. The first few (2 or 3) principal components of this physical motion analysis are compared to the first few principal components of the psychological motion analysis (as described later).

2.1 Physical Motion Analysis

The physical motion analysis consists of generating parameterized models of optical flow from image sequences. This analysis is similar to the one employed by Black et al. in their study of image motion[6]. Parameterized models can be used for motion estimation and they generally provide an accurate estimate of optical flow as many (thousands) motion dimensions (where dimensionality will be discussed later) are combined to obtain a small number of model parameters (or principal components)[7]. Starting with a set of flow fields (vectors of the optical flow of successive movie frames), PCA is used to generate a set of basis flow fields that give an approximation to the original data. These basis flow fields can later be used as a starting point for estimating optical flow in a region. This last step is not employed in this study but instead we use the derived basis flow fields to extrapolate parameters of the dance performance motion model.

A computationally expensive algorithm is used to calculate the optical flow and all of the dance data processing is performed off-line. The selected method uses image intensity data and applies a robust, course-to-fine, gradient based optical flow calculation algorithm. This algorithm is detailed in Black and Anandan's paper[7] which deals with the robust estimation of multiple motions. The actual computation of optical flow in this study utilizes Black's code and these programs can be retrieved from the internet[8]. The optical flow is calculated using a robust estimation framework that reduces the sensitivity of multiple image motion violations to the optical flow brightness constancy and the spatial smoothness assumptions. The brightness constancy constraint states that the image brightness of a region remains constant while its location changes[7]. The spatial smoothness constraint assumes that the optical flow within a region changes smoothly since it is caused by a single motion[7]. The reader should refer to Black's paper for a more detailed explanation of this optical flow technique.

To perform PCA of the dance data, the optical flow of each consecutive image is computed. Figure 15 is the computed horizontal optical flow (where white regions indicate horizontal motion and non-white regions indicate no horizontal motion) of the two right images of the three image sequence in figure 14.

Each pixel in an original image has an associated horizontal and a vertical optical flow value and these values are subsequently raster scanned (first the horizontal and then the vertical values) into a vector of length $2 \times n \times m$, where n and m are the horizontal and vertical image dimensions respectively. Each image vector forms a column of a $2 \times n \times m \times p$ matrix F, where p is the number of frames in the dance video sequence. PCA

Fig. 14: Dance image sequence

Fig. 15: Horizontal optical flow calculation

of the matrix F is then used to estimate a lower dimensional model of the motion data in F. The fimensionality of the data in the video sequence is $2 \times n \times m \times p$, because this dimensionality is very large (where n=160, m=120, and p is approximately 2500 frames), the aim is to reduce the amount of dimensions in order to make data analysis more feasible. The singular value decomposition of the matrix F can be written as:

$$F = M\Sigma V^T \tag{5}$$

where M is an $2 \times n \times m \times p$ matrix whose columns ($[\overrightarrow{m_1}, \overrightarrow{m_2}, \ldots, \overrightarrow{m_p}]$) form an orthogonal basis for F, Σ is a $p \times p$ diagonal matrix (where the diagonal values, $\lambda_1, \lambda_2, \ldots, \lambda_p$, are sorted in decreasing order), and V^T is a $p \times p$ orthogonal matrix[6]. A given flow field, $\overrightarrow{f}$, can be approximated by a linear combination of the first k basis elements of M:

$$\overrightarrow{f_k} = \sum_{i=1}^{k} a_i \overrightarrow{m_i} \tag{6}$$

where the a_i values are the principal components of the input flow field[6]. Therefore, the ith principal component, a_i, can be computed as:

$$a_i = m_i^T f_i \tag{7}$$

Figure 16 is the computed horizontal optical flow or the two right images of the three image sequence in figure 14 (and hence, an approximation of the horizontal optical flow as computed in figure 15) using the first 10 basis elements (and hence, the first 10 principal components) of the motion model matrix F.

A measure of the quality of the approximation provided by the first k columns of M is illustrated by the variance of the fraction of the matrix F accounted for by the selected (k columns) flow field components:

$$Q(k) = \left(\sum_{i=1}^{k} \lambda_i^2\right) \Big/ \left(\sum_{i=1}^{p} \lambda_i^2\right) \tag{8}$$

Fig. 16: Estimated horizontal optical flow computed from the first 10 principal components

An accurate representation of the motion model is given when $Q(k)$ approaches a value of 1.

After the orthogonal basis vectors are known for the intire dance video data set, the principal components of each frame (or input flow field) can be obtained from equation 3. Figure 17 is a sample time series (of the first three motifs, or dance video segmentations) generated by plotting the magnitude of the first two principal components at each time instance.

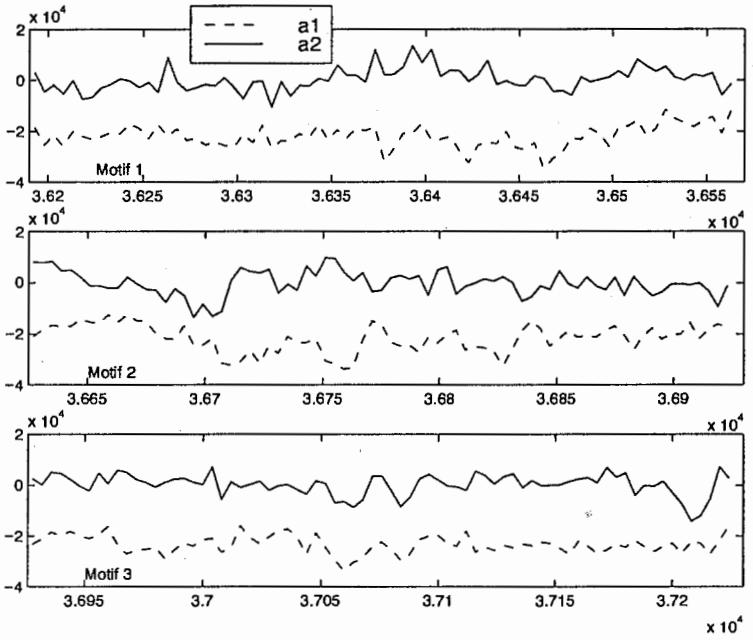

Fig. 17: Principle components of video frames

2.2 Psychological Motion Analysis

Six female subjects where asked to watch the dance and then fill out a questionaire that ranked the selected dance based on 40 different emotions in order to determine the psychological effect of the segmented dance sequence. This process was repeated twice with the same subjects. The data was then analyzed by PCA in order to narrow the dimensionality (originally a dimensionality of 40 due to the number of ranked emotion data supplied by the subjects) of the data set and in order to compare the principal components of this psychological analysis to that obtained by the physical analysis. A sample of the tabulated results of the PCA can be seen in figure 18. The 1st, 2nd, and 3rd principal components account for 44.2%, 29.1%, and 12.0% respectively of the psychological data.

Figure 19 shows the value of the first three psychological principal component magnitudes in relation to the segmented sections (or motifs) of dance. The goal is to compare this psychological principal component time series to that of the physical principal component time series.

Results of PCA Loading Factors	Principle Components			Associated Emotions
	1st PC	2nd PC	3rd PC	
1st PC (44.1%)	0.96	0.15	-0.13	noticeability
	-0.95	-0.15	0.13	complication
	-0.94	-0.10	-0.12	stability
	0.94	0.29	0.08	calmness
	*	*	*	*
	*	*	*	*
2nd PC (29.1%)	-0.01	0.97	0.17	liveliness
	-0.17	-0.93	0.15	refreshfulness
	-0.07	0.92	-0.21	brightness
	*	*	*	*
3rd PC (12.0%)	0.38	0.14	0.86	lightness
	-0.10	-0.18	0.82	depth
	-0.21	-0.53	-0.76	heaviness
Others (14.7%)	0.60	0.04	0.53	solemness
	0.57	0.17	0.44	clearness
	*	*	*	*

Fig. 18: PCA results of psychological analysis

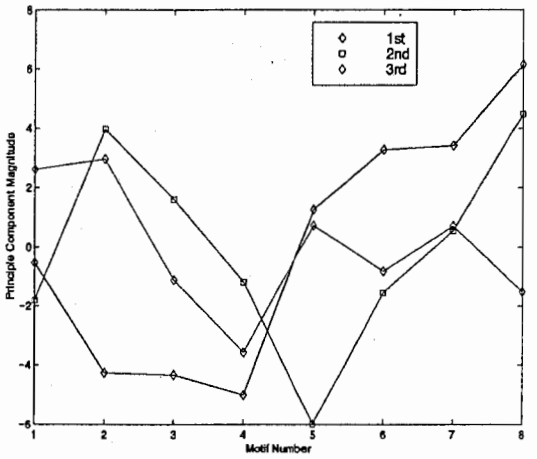

Fig. 19: Psychological impression of motif

2.3 Comparing the physical and psychological motion analysis

The goal of this analysis is to derive a correspondence between the physical and the psychological motion anaylsis. In figure 20, the magnitudes of the first two principal components for the physical and psychological motifs are compared.

The physical principal component magnitudes are generated by calculating the number of data points above the 80 and 90 percent (for the left and right plots of figure 20 respectively) of the vertical axis magnitude range line in the time series principal component data (like that of figure 17). Each motif's physical principal component magnitude (the 1st and the 2nd physical principal components respectively for the left and the right plots of figure 20) are plotted on the horizontal axis while each motif's psychological principal component magnitude (the 2nd and 1st psychological principal components respectively for the left and the the right plots of figure 20) are plotted on the vertical axis. A best fit line (in red) is drawn through the data points.

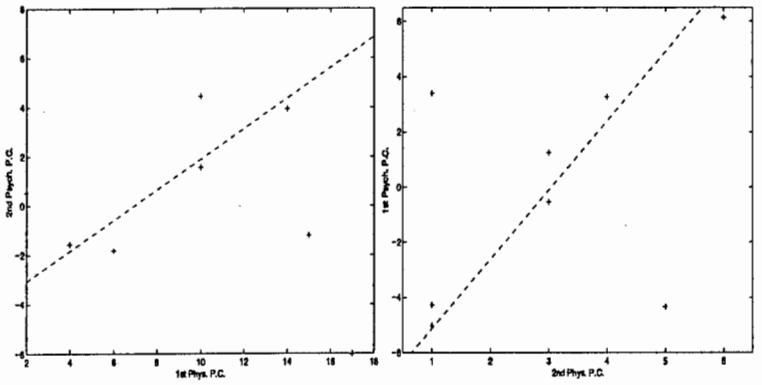

Fig. 20: Physical and psychological principal component motif correspondence

3 Dance Rhythm Estimation Preliminaries

The computer is playing a larger role in society than it ever has in the past. Part of the reason for this is that computers are becoming more accessible and user-friendly. Simple input and output nodality are no longer solely relied upon as the present computer user is presented with a wide variety of ways to more easily utilize their personal computer power.

Presented here is an example of a computer application that serves many purposes to its user. The system makes prevalent three basic concepts. First, is the idea of low-level computer based vision utilizing information from the periodicity of motion. Although this area has not been studied in detail, it is not new. The basic premise is that human motions often temporally repeat themselves in nature and hence these repetitive patterns can be used to identify the original motion patterns. Second, is the concept of a computer "teacher" that instructs a user on a given task. The proposed given task of dance instruction is developed in subsequent sections. The system presents a user, or users, with a dance composition and then provides feedback to the user as to how well they perform the dance in relation to the demonstrated dance piece. The analysis is based on the user's rhythmicity of motion and the feedback to the user comes indirectly in the form of musical alterations (as described later). Lastly, is the idea of a virtual setting where users can imagine themselves in a real-world dance setting (like a nightclub). Although the club dance setting is familiar, there is added user interaction as the users are provided with feedback (in the form of music tempo and background studio scene changes) relating to their dance "rhythm" (where "rhythm" is defined later).

3.1 Previous Work

The auditory study of rhythm with computers and music rhythm tracking has been extensively studied. Some of the more recent work on rhythm tracking has been done by Rosenthal[9] and Desain and Honing[10] where beat tracking was performed on MIDI music and also the work of Rosenthal et al.[11] and Goto and Muraoka[12] where real-time tracking of conventional audio (such as CD music), using multiple agent processing theory, was achieved.

The computer "tutor", or "teacher", idea is also not novel and such systems have been developed in the past like the work of Bobick et al.[13] and Davis and Bobick[14] for example. This first example refers to the well-known MIT KidsROOM demonstration where children are guided through a story line that is controlled, in part, by the children's actions. In the final phase of the KidsROOM, a CG instructor teaches the children how to dance. This dance is a combination of simple movements (for example, spinning) that are easily recognized by gesture recognition techniques. The second example refers to an aerobic instructor system where the user's performance is gaged on the basis of their actions as compared to those actions of a training instructor.

Dance based computer systems have been developed before and an example is the work of Paradiso and Sparacino[15]. In this system, a user's movement controls the drawing of a multi-coloured trail on a screen and also the use of various computer simulated instruments. This system provides a tool to originally generate graphics and music but it does not instructionally aide the dancer. Browstow et al.[16] are developing a system (the system is still in the development phase) proposed to act as a ballet dance instructor.

This study extends these three previously mentioned research exploits. The idea is to take the idea of a dance tutor into new genres (disco and techno musics and atmospheres) with added useability (multiple user capability for realistic dance club type settings) and different functionality (rhythm analysis versus gesture

recognition).

3.2 System

The system uses an image expression environment called the Image Expression Room[17] which consists of a virtual studio with a large display and a media handling system[17]. This virtual studio enables the chroma-keying of multi-camera input and the ability of the studio users to see themselves interacting with the system in a large half-mirror.

All software was written is C++ (with help from SGI's DMedia utilities and libraries) and it runs on a SGI HighImpact R10000 system. The display is composed of two windows which display a BIN of the dancer and also a BIN that has its intensity degregraded in time so that the motions of the dancer form a temporal template where the most recent motions have greater intensity values than those of the motions in the past (this idea is similar to the Motion Histogram Image idea that Davis and Bobick use in their motion recognition work[18]) but this feature is purely aesthetic and serves no system functionality. See figure 21 for a screen shot example.

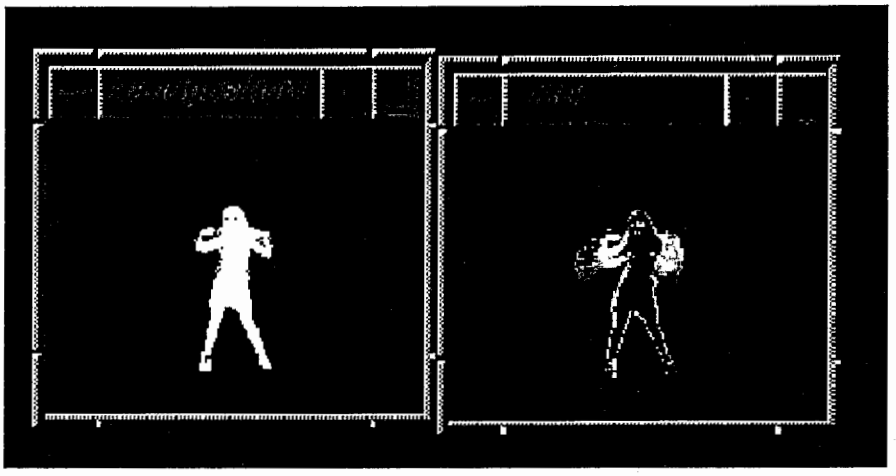

Fig. 21: System screen shot

3.2.1 System Domain

The system has two different associated music genres and MIDI formatted disco and techno music ("Stayin' Alive" by the BeeGees and "Max don't have sex with your ex" by E-Rotic) are used to guide the user through the dance sequences. The music tempo is altered (the tempo is halved if the user's rhythm is incorrect) according to the closeness of match of the user's dance rhythm to that of the system's training. Slowing the music is both an easy metric for the user to recognize and it is also a motivating factor for the user to dance back into rhythm and hence re-normalize the music tempo. The background CG (computer graphics) scene consists of multi-coloured balls that simulate the projected light off of a disco ball. See figure 22 for a scene snapshot. If the user's rhythm is incorrect, then the balls disappear from view. This background CG compositing is very simple but it is motivation for further development in this area. For instance, the atmosphere of the dance system would be greatly enhanced if there was background videos (ex. other virtual dancers visible to the user) and background images (ex. a dance club scene image) also interlaced into the system.

The word "rhythm" has a very broad meaning and hence the classical view of musical rhythm is slightly different than the one used in this study. Rhythm is assumed to mean measured movement, as in dancing[19]. The goal is to estimate a user's dance rhythm, or measured movement, as compared to that of a skilled dancer rhythm. With this in mind, a very simple method of the frequency analysis of measured movement is proposed. At each time instant, a BIN is calculated for each pair of adjacent frames. If there is motion in a pixel, then this fact will be reflected in the BIN and the number of "motion pixels" are counted at each time instant. Therefore, there is an associated number of pixels that experience motion for each pair of consecutive frames. This associated number is inserted into a "windowed" vector by a queueing (the newest value enters and the

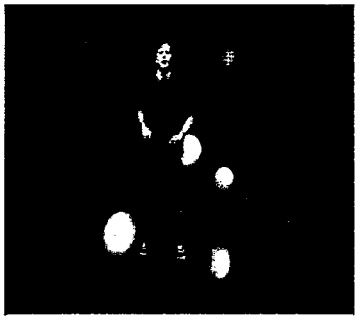

Fig. 22: System screen shot

oldest value exits) data structure and therefore the vector is kept at a constant length. This length is set at 300 values as this is approximately the minimum amount of data needed to get an accurate dance frequency representation for the particular training data used. At each time instant, the MEM transform is taken of this vector in order to map the windowed time series into the frequency domain. A frequency difference measure (as described later) is calculated by comparing the frequency of the characteristic peaks of the live video sequence to those of the training dance sequence.

One advantage of this simple rhythm estimation method is that the relative magnitudes of motion have no effect on the analysis (of course the power spectrum magnitude changes but the respective characteristic peak frequencies are the same). Instead, the rhythm is formed by changes in the dancer's motion that result in changes of the motion time series. For example, if the dancer's hips are swaying back and forth, then the time that a left movement direction changes to a right movement direction results in a timeseries magnitude change (in this case, a minimum value). For the image sequence illustrated by figure 23 one can see the resulting time series in figure 24. Note that the graph minima correspond to direction changes in the dancer's motion.

Fig. 23: Frames 2475, 2485, and 2497 in motion direction change example

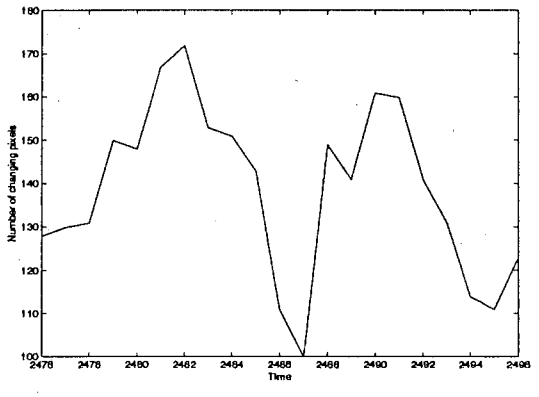

Fig. 24: Motion time series for image sequence represented by figure 4

Hence, the proportions (body size, clothing styles, ect.) of the live dancer and those of the training sequence dancer need not be the same nor normalized. This also enables the possibility of multi-user functionality. If there is two people in the studio and they are both moving in rhythm as compared to the training dancer, then there is no difference (in terms of frequency analysis) than if there is only one person in the studio dancing in rhythm.

There may be a more robust way to estimate the dancer's motion or motion direction change. For example, it may be possible to use the MHI (as described in section 2.1) to calculate the times of motion direction change. This time would correspond to the maxima of the derivative of image motion change. It is not immediately apparent how this analysis could be carried out.

3.2.2 Training Data

The system is first trained with the rhythm of a skilled dancer in order to provide feedback on a dancer's rhythm. Ekaterina Saenko (from the University of British Columbia) provided the dance sequences for the training disco and techno dance style data. Ekaterina has studied dance for over 10 years and she is very skilled in both classical and modern dance. She performed two sequences of typical disco and techno dance styles. Underlying both performances is a constant body rhythm (usually achieved by swaying hip movement) and relatively simple arm and leg movements. See figures 25 and 26 respectively for examples of disco and techno dance rhythm sequences.

Fig. 25: Examples of disco dance sequences

Fig. 26: Examples of techno dance sequences

The complete time series (whole video) of motion change values is transformed into the frequency domain and the peak frequency values are calculated. These values are used in the live studio rhythm processing system. Figure 27 shows the characteristic peaks of the disco and techno dance sequences.

3.3 Frequency Analysis

Although the Fast Fourier transform (FFT) is a frequently used tool for frequency analysis, other methods of analysis often prove to be more useful than the FFT[20]. In this study we chose to use a method called the Maximum Entropy Method (MEM). The FFT (figure 28) was used to do a similar analysis to that of figure 27.

The frequency peaks are harder to identify in figure 28 than those in figure 27. The FFT power spectrum of any real-valued time series function, $c(t_k)$, is represented by:

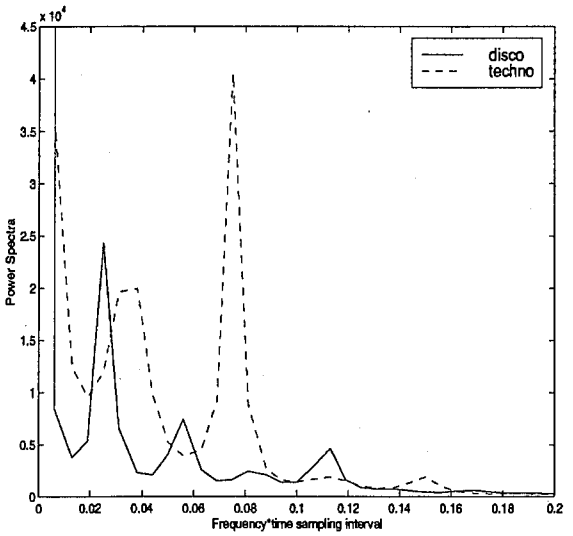

Fig. 27: Computed frequencies for disco and techno musics

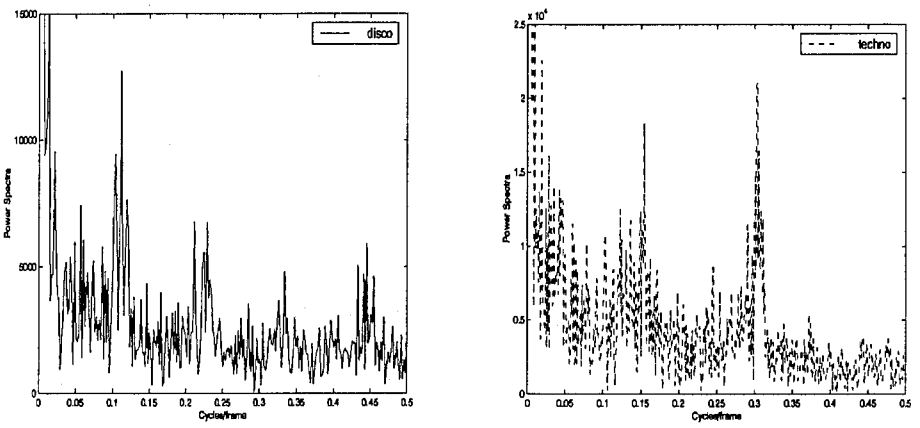

Fig. 28: FFT computed frequencies for disco and techno musics

$$P(f) = \left| \sum_{k=-\infty}^{\infty} c(t_k) e^{2\pi i k f \Delta} \right|^2 \tag{9}$$

where f contains the frequencies in the Nyquist range, $-f_c < f < f_c$, and Δ is the time domain sampling interval. A common approximation to equation 1 is:

$$P(f) = \left| \sum_{k=-N/2}^{N/2} c(t_k) e^{2\pi i k f \Delta} \right|^2 \tag{10}$$

but it often turns out that equation 1 is approximated better by:

$$P(f) = \frac{a_0}{\left| 1 + \sum_{k=1}^{M} a_k e^{2\pi i k f \Delta} \right|^2} \tag{11}$$

Equation 3 is called the MEM and to solve it, the values of a_k, $k = 0, 1, \ldots, M$ need to be found. The difference between equations 2 and 3 is that equation 3 can have *poles* which provide accuracy when the power spectrum has sharp peaks. Equation 2 can only approximate the spectral peaks of equation 1 by finding the zeroes of the polynomial and not the poles[20].

Linear prediction theory provides the values of a_0 and a_k. In classic linear prediction, one wants to find the next value of M consecutively linearly spaced points of data y_i. The linear prediction formula is:

$$y_n = \sum_{j=1}^{M} d_j y_{n-j} + x_n \tag{12}$$

where y_n is the magnitude of the data point x_n and the linear prediction (LP) coefficients, d_j, are found using autocorrelation:

$$\phi_j \approx \frac{1}{N-j} \sum_{i=1}^{N-j} y_i y_{i+j} \tag{13}$$

and:

$$\phi_k = \sum_{j=1}^{M} \phi_{|j-k|} d_j \tag{14}$$

where $k = 1, \ldots, M$. The values of the a_k values in equation 3 are obtained from the LP coefficients:

$$a_0 = xms \qquad a_k = -d_k \tag{15}$$

where xms is:

$$\langle x_n^2 \rangle \equiv xms = \phi_0 - \phi_1 d_1 - \phi_2 d_2 - \ldots - \phi_M d_M \tag{16}$$

3.3.1 Frequency Distance Measure

The closeness of fit between the training and the test data power spectrum vector is measured once the power spectrum of the windowed time series data is computed. More specifically, the goal is to measure the similarity of the training and test characteristic spectral peak frequencies and hence the use of various normality tests, like correlation or the Kolmogorov-Smirnov test which measure the similarities of the whole data set, are not adequate. Therefore, a specialized power spectrum vector distance measure is used.

The characteristic peak frequency values ($cpfv(j)$ where $j = 1, \ldots, p$ and p is the number of characteristic peaks of the training data) of the training data are calculated before run-time. The test data power spectrum vector is partitioned into n (the optimal value of n is determined during the training phase but this could be done automatically during the test phase run-time) segments during execution. The corresponding frequency value of the maximum power spectra of the test data for each segment is saved in a vector, $datapeakfreqs(i)$ (where $i = 1, \ldots, n$). For each value of $cpfv(j)$, the vector $datapeakfreqs(i)$ is searched for the smallest difference, $divdiff(j)$, between the particular value of $cpfv(j)$ and each element of $datapeakfreqs(i)$. All the values of $divdiff(j)$ are added to give a total difference measure, $totdiff$, for all the characteristic peaks of the test data as compared to the those of the training data. Pseudo-code for this distance measure technique is indicated in figure 9:

```
initialize totdiff and divdiff(j) to 0
segment test data power spectrum into n parts
for  i = 1 to n,
     save frequency of the maximum power spectrum value of segment i in datapeakfreqs(i)
end
for  j = 1 to p,
     divdiff(j) = maximum floating-point number
     for  k = 1 to n,
          if |datapeaks(k) - cpfv(j)| < divdiff(j)
            divdiff(j) = |datapeaks(k) - cpfv(j)|
          end
     end
     totdiff = totdiff + divdiff(j)
end
```

Fig. 29: Pseudo-code for distance measure

3.4 Problems and Improvements

The system described here is still work in progress and hence there are several inefficiencies. The first deals with ambiguities such as what exactly is the rhythm of dance and what types of dance the choice of this definition applies to? It is agreed that the definition of image rhythm adopted here is not the only one available. This type of rhythm analysis is dance specific. The styles of dance are so numerous that it is uncertain whether or not a unified dance rhythm analysis method is possible or even sensical.

The rhythm analysis algorithm computes in real-time (at about 30 Hz) but the actual response of the system to the dancer's motion has a delay associated with it. This delay corresponds to the time for a data value to propagate through the windowed time series of pixel motion values. The system would be more effective if this delay was negligible. Furthermore, in theory it should be possible to obtain more rhythm information than a single distance measure value. With a small change in the distance measure algorithm it would be possible to differentiate frequency values that are greater or smaller than those of the training data. This would make it possible to provide more detailed feedback (decrease or increase the speed of the music) to the user but in reality, the algorithm is not precise enough to do this.

The training data is very specific and hence the system works best when the user imitates this dance style exactly. Ekaterina maintains a fairly constant rhythm throughout her dancing although there are periods when changes in rhythm happen (ex. when she tires and stops or slows down momentarily). The training data would be more robust if a more strictly controlled dance sequence is used.

4 Conclusion

Three analysis techniques for video rhythm and motion analysis are presented. First, spatial frequency analysis is proposed to determine the composition of an image or series of images. With this method it is demonstrated that it is possible to distinguish between two different movies. It is also shown that close-up and long camera shots can be recognized in ideal conditions. Second, dance motion is analyzed by two contrasting methods. The motion is first analyzed physically by an optical flow and principal component analysis technique. Next, the motion is analyzed by a subjective psychological experiment and principal component analysis. The resulting physical and psychological principal components are compared in order to derive a relationship between people's emotionals as produced by dance and physical analysis. A relationship showing the correspondence between the physical and psychological analysis seems to be evident. Lastly, a real-time system used for the estimation of dance rhythm is described. The user is guided through a dance sequence and subsequently receives feedback as to their dance composition as compared to that of a trained dancer. The user receives indirect feedback in the form of audio (musical tempo change) and visual (the presence or lack of a CG background). The system estimates the user's dance rhythm by using BIN pixel change information and a windowed MEM frequency method.

5 Acknowledgements

I would like to thank those who helped with this work. Most notably, a big thank you to Ekaterina Saenko for the open house training dance sequences and all of department 3 in the ATR Multimedia Integration and Communications Research Lab for their help.

References

[1] R. Suzuki: *Personal Communications*, 1998.

[2] J. Evans and G. P. Morriss: *Digital Pictures: Representation and Compression*, New York: Plenum Press, 1998.

[3] A. Netravali and B. Haskell: *Digital Pictures*, Orlando: Academic Press, 1982.

[4] A. Rosenfield and A. Kak: *Digital Picture Processing*, Orlando: Academic Press, Vol.1, 10, 1982.

[5] K.H. Pribram: *Brain and Perception*, New Jersey: Lawrence Erlbaum Associates, 1991.

[6] M.J. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. *CVPR'97*, 1997.

[7] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75-104, January 1996.

[8] http://www.parc.xerox.com/spl/members/black/regression.html

[9] D. Rosenthal. *Machine rhythm: Computer emulation of human rhythm perception*, Ph.D. Thesis, Massachusetts Institute of Technology, 1992.

[10] P. Desain and H. Honing. Quantization of musical time: A connectionist approach. *Computer Music Journal*, 13:56-66, 1989.

[11] D. Rosenthal, M. Goto, and Y. Muraoka. Rhythm tracking using multiple hypotheses. *Proc. of the 1994 Intl. Computer Music Conf*, pp.85-87, 1994.

[12] M. Goto and Y. Muraoka. Music understanding at the beat level - Real-time beat tracking for audio signals. *IJCAI-95 Workshop on Computational Auditory Scene Analysis*, pp.68-75, August 1995.

[13] A. Bobick, J. Davis, S. Intille, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. Kidsroom: Action recognition in an interactive story environment. PerCom TR 398, MIT Media Lab, 1996.

[14] J. Davis and A. Bobick. Virtual PAT: A virtual personal aerobics trainer. PerCom TR 436, MIT Media Lab, 1998.

[15] J.A. Paradiso and F. Sparacino. Optical tracking for music and dance performance. *Fourth Conference on Optical 3-D Measurement Techniques*, September 1997.

[16] http://www.cc.gatech.edu/gvu/perception/projects/ballet/index.html

[17] S. Inoue, M. Ishiwaka, S. Tanaka, and J.I. Park. An image expression room. *Proc. Intl' Conf. on Virtual Systems and Multimedia'97*, pp.178-186, September 1997.

[18] J. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. *Proc. Comp. Vis. and Pattern Rec.*, pp.928-934, June 1997.

[19] *The Random House Dictionary of the English Language: 2nd Edition.* New York: Random House, 1987.

[20] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing Second Edition.* Cambridge University Press, 1992.