Dancing with Turks (ACM Multimedia 2015, long paper)

Conference Paper · October 2015

DOI: 10.13140/RG.2.1.2692.3609

6 authors, including:

Yanxi Liu

Pennsylvania State University

163 PUBLICATIONS **4,429** CITATIONS

SEE PROFILE

Dancing with *Turks*

I-Kao Chiang[◇] Ian Spiro[†] Seungkyu Lee[⋆] Alyssa Lees[†] Jingchen Liu[‡]

Chris Bregler[†] Yanxi Liu[‡]

[◇]Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
[†]Dept. of Computer Science, Courant Institute, New York University, New York, NY, USA
[⋆]Dept. of Computer Engineering, KyungHee University, Yongin-si, South Korea
[‡]School of Electrical Engineering and Computer Science,
The Pennsylvania State University, University Park, PA, USA
{igorchiang,ispiro,seungkyu74,alyssa.lees,jasonliu401,chris.bregler}@gmail.com
yanxi@cse.psu.edu (corresponding author)

ABSTRACT

Dance is a dynamic art form that reflects a wide range of cultural diversity and individuality. With the advancement of motion-capture technology combined with crowd-sourcing and machine learning algorithms, we explore the complex relationship between perceived dance quality/dancer's gender and dance movements/music respectively. As a feasibility study, we construct a computational framework for an analysis-synthesis-feedback loop using a novel multimedia *dance-music texture* representation. Furthermore, we integrate crowd-sourcing, music and motion-capture data, and machine learning-based methods for dance segmentation, analysis and synthesis of new dancers. A quantitative validation of this framework on a motion-capture dataset of 172 dancers evaluated by more than 400 independent on-line raters demonstrates significant correlation between human perception and the algorithmically intended dance quality or gender of synthesized dancers. The technology illustrated in this work has a high potential to advance the multimedia entertainment industry via *dancing with Turks*.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Learning*

Keywords

Dance perception; Crowd-sourcing; Motion-capture; Dance-texture; Feature selection; Regression; Dance-texture synthesis; Animation

1. INTRODUCTION

As digital media comes of age, massive multi-media data sets have become commonplace while computational tools to discover their intricate relations are scarce. With the introduction of crowd-sourcing technology, human perception-driven algorithms become more feasible. In social (non-choreographed) dance, perception by others plays an important role to our understanding of social dynamics among dancers and viewers. Our goal in this work is to develop a computational framework and associated tools for dance perception, driven by active, on-line audience participation. The feasibility of our framework has been validated on a multi-media, multi-dimensional data set composed of motion capture data of 172 Jamaican dancers of both genders free-dancing to the same music. Our initial investigation focuses on the dance quality (good, bad) and dancer's gender (masculine, feminine), since the perception of these two attributes has a direct impact on social dynamics reflected in social dance across different cultures.

We combine a suite of signal processing, machine learning, animation and social media technology to discover mappings between human dance perception and a set of measurable, isolatable and controllable human body joint parameters. More specifically, from an initial set of 25 human ratings per dancer, we use a weakly supervised learning method to learn dance ability and dancer's gender discriminative features composed of subsets of body joints (angular displacement, velocity and acceleration) and to construct a multivariate regression model that captures human-perceived dance quality and dancer's gender. We then employ these learned features as "control knobs" for music-motion-correlated dance segments to synthesize ability or gender-targeted new dancers. To close the loop, the synthesized dancers are re-evaluated via crowd-sourcing by hundreds of online Mechanical Turk workers (Figure 1). The main contributions of our work:
(1) a novel machine learning-based, human perception-driven computational framework for dance analysis, synthesis and validation from multimedia data (music, motion-capture data and human ratings);
(2) a novel spatiotemporal *dance texture* and *music texture* joint representation, and a novel dance texture synthesis algorithm, confirmed to be visually convincing by Mechanical Turks with statistical significance;
(3) an effective human-in-the-loop initiative that uses crowd-sourcing both as an initialization for machine learning and as a validation for the dance analysis and synthesis algorithms (Figure 1).

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806220.

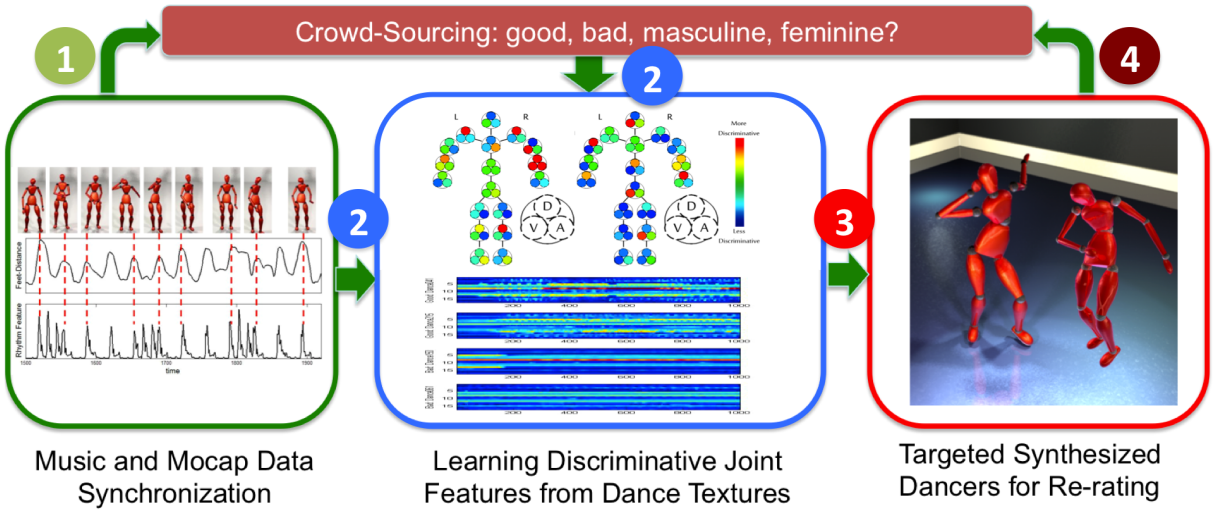


Figure 1: The outline of our multi-media computational framework: after computing the synchronization between the raw mocap and music data (Section 3.1) we (1) obtain a set of initial human ratings via crowd-sourcing on dance ability and dancer's gender; (2) input both synchronized mocap/music data and the human ratings for analysis (Section 3.2), where we discover, automatically, the most discriminative human joint feature subsets (D: displacement, V: velocity, A: acceleration) for dance ability and dancer's gender respectively; (3) synthesize targeted new dancers (for example: a level-5 (highest level) dancer or a female dancer) using learned discriminative body joint features (Section 3.3); and (4) feed the newly synthesized targeted dancers to on-line human raters (the Mechanical Turks) for re-evaluation of dance ability and dancer's gender (Section 3.4). Finally, we compare the statistical consistency between human perception and the intended/targeted dancer's ability and gender (Table 2).

2. RELATED WORK

"Human in the loop" applications span many research disciplines from the creation of simulation software like flight simulators to the more active solicitations of human feedback for image retrieval [37, 27]. Using crowd-sourcing to improve the performance of computer vision algorithms has become more feasible in recent years, especially for those data sets with visually distinct objects, such as identification of birds, and parts of a bird[1]. Many research disciplines such as machine learning, mathematical optimization, automatic control, cyber-physical systems, and autonomic computing rely on feedback to achieve goals such as autonomy, learning, adaptation, stabilization, robustness, or performance optimization[2]. Our effort differs from these in that the human perception of dance we use is at a *general impression* level instead of specific object instance/parts or feature labeling by the human users. The localization details (which body joints) are learned automatically in our work. Many *motion perceptual experiments* are relevant to our work, dating back to the famous Johannsson Experiments [14], particularly those perceptual experiments on gender classification [16, 30]. There is also a vast literature on automatic motion analysis techniques [8] beyond the scope of this paper.

Previous work in style transfer (e.g. transforming from "happy" to "sad" while preserving walking to location A) has looked at applying user-specified style to an existing motion sequence [34, 13, 32, 7, 28], or existing style in a motion capture database to new motion paths [12, 1, 15, 17, 26, 33, 36]. Most of the recent techniques build a graph structure out of large amounts of motion capture data [1, 15] and have been inspired by earlier systems in speech recognition [23], video face animations [4], and video textures [29]. Related to style transfer are methods that integrate stylistic variables into generative models to synthesize stylized motion directly. Brand and Hertzmann [3] modeled style and content using HMMs whose emission distributions depend on stylistic parameters learned from data. More sophisticated models have been introduced using other HMM extensions, Gaussian Processes and hierarchical models applied to style variations [18, 35, 25]. The most complex architecture, based on so called "deep networks", has been introduced by Taylor and Hinton using explicit stylistic variables to modulate the interactions of binary latent variables modeling dynamics and real-valued variables representing pose [31].

Different from example-based motion synthesis, our work depends on fuzzy input from human raters that may or may not agree with each other. We face both the question of how to use such inconsistent labeling as well as figuring out what body motion attributes dominate human perceptual evaluation.

Directed graphs have been used, e.g. [15], as a primary motion/dance representation with edges corresponding to clips of motion and nodes as choice of connecting points. Other dance representations are typically one dimensional along the time axis [18, 26]. Our *dance texture* proposed in this work is a true 2-dimensional image (Section 3.1). Our dance texture synthesis algorithm is the first to adapt the "*Markov properties*", employed in image-based near-regular texture synthesis [11, 19, 22], into a multimodal audio-spatio-temporal *dance texture* space.

[1] http://www.vision.caltech.edu/visipedia/
[2] *The Feedback Computing workshop 2013*

3. OUR APPROACH

We start with a given piece of music, *Elephant Man*, sampled at a standard 44.1 kHz stereo, and a set of *asynchronized* motion capture (mocap) data of 172 Jamaican (teenage) dancers of both genders (equal distribution) free-dancing to the same music[3]. Independently, we collected 25 human ratings for each dancer on both the perceived dance-ability and the perceived gender, where each human rater is shown a replay of a dancer's mocap data on a computer screen, and asked to provide a score between 0 (female) and 1 (male) for gender; and a score between 1 (bad) and 5 (good) for dance quality or dancer's ability. We use the average of ability (gender) ratings for each dancer in our subsequent analysis.

Our goal is to develop computational means that can (1) learn the most discriminative body-joint features for regression models best corresponding to human ratings; (2) use this knowledge to generate new dancers with intended dance ability or gender; and (3) validate the synthesized dancers via human perception using independent human raters (Mechanical Turks).

To facilitate the discovery of the intricate relationships between human perception and the body movements, we first propose a spatiotemporal representation of mocap data, *Dance Texture* (Figure 2), and synchronize the 2-dimensional dance texture with the 1-dimensional music texture along the time axis.

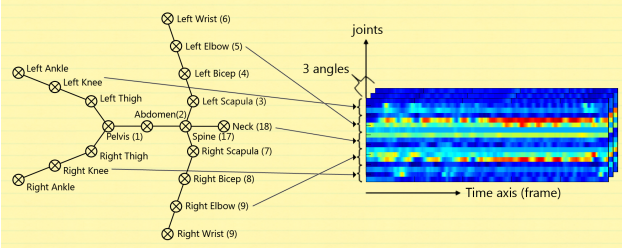


Figure 2: Left: The 18 body joints captured by the motion capture data. Right: A dance texture of three layers, $P_{axis}(time, joints)$, $axis : X, Y, Z$, capturing joint rotation angles about each axis. P is a 2D frieze-like pattern/image, extending along the (horizontal) time axis bilaterally centered about the four mid-body-joints: neck, spine, abdomen, and pelvis.

3.1 Dance Texture and Music Texture

Dance Texture

Since human motion capture (mocap) data captures joint rotation angles about the corresponding joint axes X, Y, Z respectively [5, 24], we define a dance texture $P_{axis}(time, joints)$, where $axis : X, Y, Z$. Each P_{axis} is 2D frieze-like pattern extending along the time axis horizontally from left to right (Figure 2). Since the human body presents a natural bilateral symmetry, we place the left-right body joints correspondingly around the four mid-body joints. Each point on the dance texture P_{axis} denotes the twist angle value about its rotation axis $X(Y, Z)$, represented here in the color scheme of $[R, G, B]$. Each column of $P_{axis}(time = i, :)$ represents a pose of the dancer at time i and each row of $P_{axis}(:, joint = j)$ represents the angular values of a specific body joint j along time axis. For simplicity, we sometimes omit the rotation axis indicator X, Y, Z in $P_{axis}(i, j)$ by summing by up all twist angles at i,j as $P(i, j)$.

Music Texture

When transforming music input into the frequency domain using conventional Fourier transform, the frequencies corresponding to integer multiples of 0.44 Hz and 0.88 Hz i.e. harmonics of 0.44 Hz, stand out in the music power spectrum and the average dancer power spectrum respectively (Fig.3 (A)). This clearly suggests that the music and the dancers have commensurate periodicity and the ratio between the fundamental frequencies is 2 (0.88 Hz/0.44 Hz).

Dance Texture and Music Texture Synchronization

The corresponding periods in the dance and music textures are captured computationally, while the alignment between dance and music remains ambiguous. Exploring different body joints, a strong correlation between the feet separation and the up-beat is observed and captured computationally. Since the music and body-motion signals have shared cycle lengths, we only need to compute the offset within a single period T. The beat signal $b(t)$ is obtained by passing the music signal through an anti-aliasing low pass filter and then downsampling to the same $30Hz$ sampling rate as the mocap data. The music beat signal (down sampled), between-feet distance and their correlations are displayed in Figure 3 (B).

3.2 Analysis

To further explore the discriminative power of body joints in relation to human ratings, we define and extract measurements from dance texture beyond angular joint values.

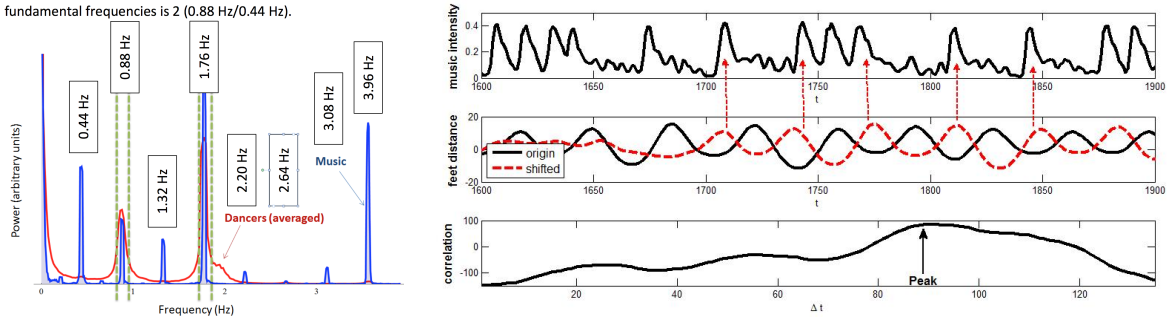
Dance Texture Features

We extract from three types of dance textures: angular Displacement (D), angular Velocity (V), and angular Acceleration (A). Given a dance texture $P^d(i, j)$ for a specific dancer d with i indicating the frame number (time) and j body joint, the total number of raw features is 3 (D,V,A) $\times$18 (body joints) $\times$# frames = 54 $\times$ # frames (Figure 2). One approximation of these features is an accumulated movement of each joint $j \in [1...18]$ over the entire dance segment: $P^d_\Sigma(j) = \sum_{i=1}^{n} P^d(i, j)$ and all the possible subsets of 18, from a single joint to nine joints (and their complements), to form our initial feature set $F^d_{Displacement}$ short as F^d_D: $|F_D| = \sum_{j=1}^{9} \binom{18}{j} = 155,381$. Including angular displacements (D), angular velocity (V) and acceleration (A) measurements in an analogous manner, we obtain a feature set F^d_{DVA}, short as F^d with $|F^d| = 155,381 \times 3 = 466,143$. Since every dance has the same set of features, we can drop d and use F to express this total feature set from dancer mocap data of 172 data points (dancers) where each data point has 466,143 feature dimensions.

Dimensionality Reduction for Regression

Given the distribution of human ratings (green circles in Figure 6) for dance quality and dancers gender respectively, we propose to construct a linear regression model to approximate human perceptual labels. It is obvious that some dimensionality reduction of the dance-texture feature set F is necessary for us to learn and compute a model capturing the most relevant parameters of human perception of dance. We have experimented with multiple methods for dimensionality reduction in this work. We can summarize

[3]We obtained the raw motion capture dataset from [6] where it was used for human visualization only.



(A) (B)

Figure 3: (A) Using the conventional Fourier transform spectral method to show representative power spectra of the music and the mocap of the dancers. The red line corresponds to the intensity averaged across all 172 dancers. The green dotted lines indicate the two regions where the integrated intensity is correlated to dancer ability. (B) Alignment of the music (30Hz sampling rate, top), between-feet distance (30Hz sampling rate, middle), and correlation score (bottom).

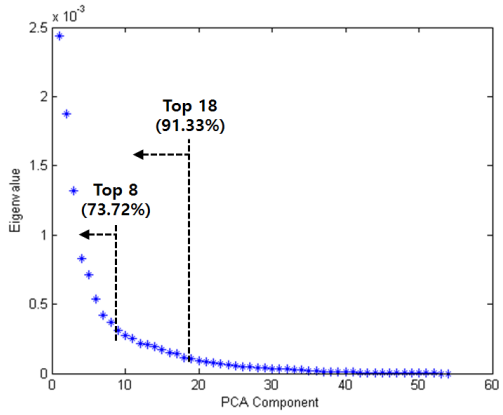


Figure 4: Eigenvalues versus Eigenvectors plot from a principle component analysis (PCA) of the dancers mocap data set. For dimensionality reduction, we experimented with taking the top 8 and top 18 eigenvectors to construct two corresponding linear regression models (Figure 6).

them into two basic categories: Principle Component Analysis (PCA)-based and machine learning-based methods.

PCA-based: For dimensionality reduction, we apply Principle Component Analysis on the raw mocap features of angular displacements for all subjects. Figure 4 shows the eigenvalue versus eigenvector plot. Using the top-8 (73% variance) and top-18 (91.33% variance) eigenvectors respectively, we built and evaluated two linear regression models against human ratings (Tabel 1, Figure 6). We pick top-8 eigenvectors since previous works using PCA methods report 8 principle components (PCs) as sufficient for capturing human motions such like gaits; and top-18 is due to the fact that 18 different body joints are captured in the dance mocap data set (Figure 2). It is worth noting that PCA is an unsupervised approach for data analysis, thus the human ratings associated with the data set play no role in the construction of the PCA-based linear regression models.

Machine Learning-based: Supervised learning is an active branch of research in general machine learning to automatically find the most discriminative features that separate different classes of data given the class labels [2]. In our case, perceptual ratings of multiple human raters are readily available, the average ratings are however not necessary discrete 'class labels' to train a classifier. We have experimented with two discriminative feature subset selection measures, which are used to select the highest ranked features to construct linear regression models for predicting human ratings on dance quality and dancer's gender respectively, achieving dimensionality reduction at the same time.

Table 1: Comparisons of three regression models measured by Root-Mean-Squares (RMS) against human ratings, the RMS value is the smaller the better. Also see Figure 6.

PCA (variance captured)	Ability	Gender
Top-8 Eigenvectors (73.72%)	0.0077	0.0321
Top-18 Eigenvectors (91.33%)	0.0048	0.0218
Proposed Method	**0.0009**	**0.0087**

Learning-based Method #1: The first method is to find a continuous correlation value between human ratings and feature values in the dance-texture feature set F. For a feature of dancer d, $f^d \in F$, its relation to perceived rating R^d_{gender} of dancer d can be expressed as: $f = a \times R_{gender} + b$, and solved for a, b using all dancers. A $(R, p\text{-}value)$ pair can thus be computed for all features $f \in F$ and used to rank all features by their correlation value R to the perceived dancer ability or gender ratings respectively. We then pick the top N features as the most discriminative features for gender (dance ability) separation and build a pair of linear regression models that are quantitatively compared with the PCA-based regression models above (Table 1, Figure 6).

Learning-based Method #2: Alternatively, we train a set of binary-class classifiers by modifying the given human ratings. We sample a data subset where only the dancers with human ratings in the extreme ranges (say thresholds at ratings $> 75\%$ as class 1 and $\leq 25\%$ as class 0) are chosen, and use them to form a binary classification problem with discrete class labels.

We use an *augmented variance ratio* (AVR) as a criteria to compute and rank discriminative features for these two classification problems (ability and gender) respectively. AVR is a variant of Fisher criterion [10] and has been used in many other biomedical image applications, e.g. [20, 21]. To make the features more compact, we perform PCA on the top N most discriminative features. A multivariate linear regression model is then constructed using the PCs found above. This regression model is trained and cross-validated using leave-one-out (LOO) on the same dance segments (500 frames/segment) the *MTurkers* have rated. We have obtained the average error rates of $12 \pm 8.6\%$ for dance ability and $14.58 \pm 10.66\%$ for gender.

The machine learning-based methods have helped us to gain new insights on the human perception of dancers:
(1) *Which body features are important?*
Figure 5 illustrates three different feature selection results demonstrating, qualitatively and quantitatively, which body joints and what type of joint measures (Displacement, Velocity, Acceleration) are discriminative for human perception of dancer quality and/or dancer's gender.
(2) *How to label/rate a given dancer algorithmically that mimics a human rater?*
The regression models built from the selected discriminative features provide a higher level, dance quality/dancer's gender *predictor* that maps an arbitrary length of dance (mocap data) into a pair of perceptual labels. This function will be used extensively in our new dance synthesis algorithm next.

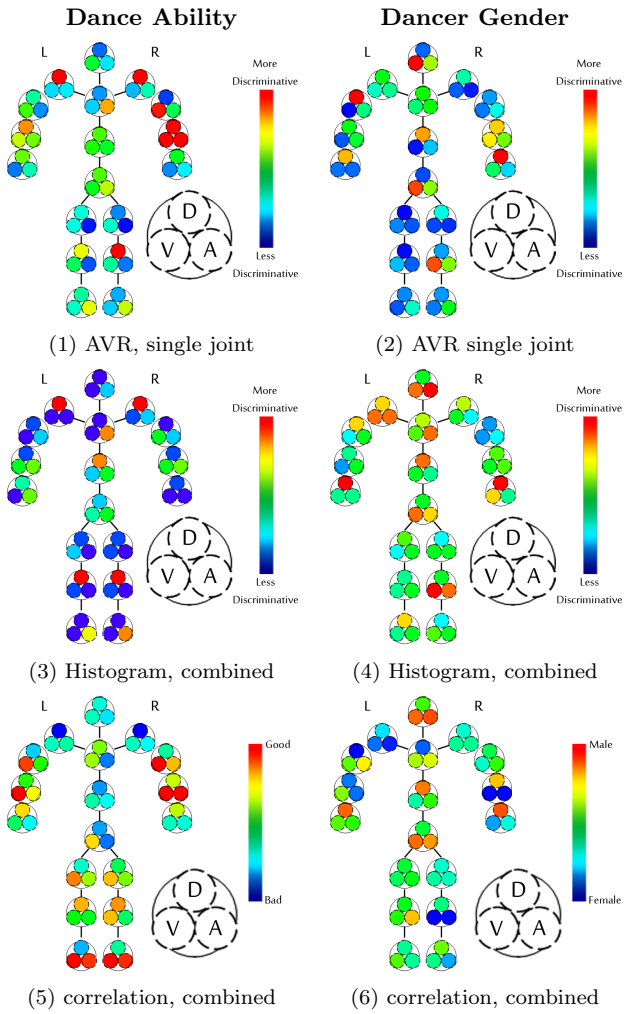
Automatic Dance Texel Segmentation and Rating

As a result of music texture and dance texture synchronization (Section 3.1), we are able to segment a dance texture into a sequence of equal sized **dance-texture texels** along the time axis. Dance-texture texels are the smallest dance-texture units, each of which has the length of one music beat/measure (Figure 3), that can also be used as the building blocks for synthesizing new dancers. Since each dance texel is a short dance (one measure long), we can map it to the discriminative feature space and use our machine learning-based linear regression model to *rate* each dance texel on its dance quality and its dancer's gender. As a result, we have a collection of dance texels that are unit length and are automatically labeled in both dance quality and dancer's gender.

This computational *decomposition* of a long dance texture into small pieces (texels) of short dances, and the semantic labeling of each dance texel play a crucial role in: (1) Perception – each dance/dancer can be viewed as a composition of many dance texels; each texel can have its independent labels in dance quality and dancer's gender. Thus, during a long dance, it is possible to perceive a dancer as sometimes good and sometimes bad, sometimes more masculine and sometimes more feminine; (2) Synthesis – these dance texels, associated with their own ratings from our learned regression model, form the perfect dance texture synthesis *building blocks* to generate new dancers with the desired perceptual effect. Therefore, the dance-synthesis process with desired perceived gender and ability becomes a dance-texel composition procedure of mixing and matching texels from **different** dancers selectively.

3.3 Synthesis

A unique aspect of our approach is to treat dance as a 2D texture, enabling the generation of a new dance via tex-



(1) AVR, single joint (2) AVR single joint

(3) Histogram, combined (4) Histogram, combined

(5) correlation, combined (6) correlation, combined

Figure 5: Using D (displacement), V (velocity), and A (acceleration) body joint features, we demonstrate the most discriminative features learned algorithmically using three different computable measures. (1, 2) are the single joint features (D,V,A) ranked and color-coded (red: most discriminative) by its respective Augmented Variance Ratio (AVR). (3,4) are combined-joint features evaluated using the histogram of the most selected (highly frequent) body joint features (from ten random splits of data into training/testing subsets). (5,6) are the feature ranking results from the correlation between *feature-subset* (D,V,A) values and the human raters labels. Several corresponding body joint features are found across three different discriminative feature selection measures. For example, all three find that left and right knee displacement and ankle velocity/acceleration in particular are most discriminative for dance quality; while the right (and left) wrist joint angel displacement, the head/neck angle velocity/acceleration, abdomen angle displacement, and pelvis angle velocity/acceleration (mostly upper body joints) highly discriminative for dancer's gender perception. Interestingly, many of these automatically found, discriminative joint features are also acknowledged by some human raters (Table 2).

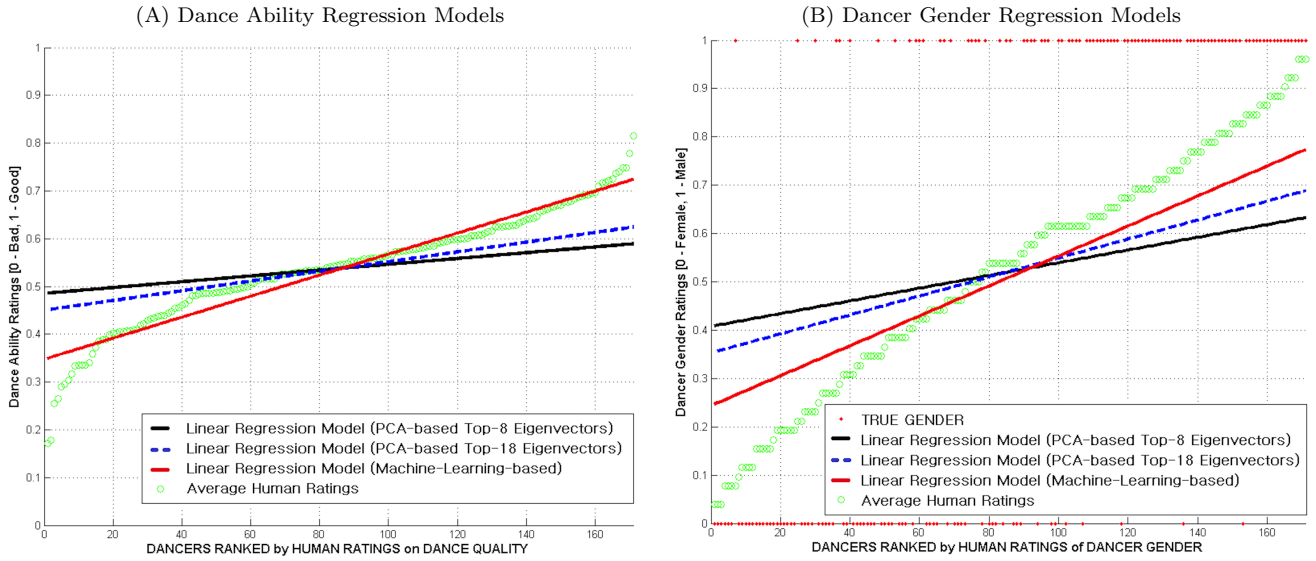


Figure 6: We compare our proposed supervised-learning-based linear regression model against non-supervised PCA-based dimensionality reduction linear models, where different top-N eigenvectors with the highest eigenvalues are used (top-8, top-18). The horizontal axes above are ranked dancers in non-decreasing order of their average human ratings (25) of dancer ability and dancer's gender respectively: bad dancer=0, good dancer=1; female=0, male=1. The three overlaid linear regression model plots, two from PCA-method using top-8 and top-18 eigenvectors respectively and one from our feature-selection method, illustrate that our proposed machine learning-based regression models align the best with human perception of dancers in terms of their dance ability and demonstrated gender; while our dance ability regression model has a better fit with human ratings than the regression model for dancers genders. Table 1 shows the root-mean-square (RMS) measures between these regression models and human ratings. It is worth noting that (1) the average human ratings (green circles) are not strictly linear either for dance quality (left) or for dancer's gender (right); and (2) the average accuracy of human gender rating with respect to dancers' true genders (red dots) is only $64.78\% \pm 7.6\%$.

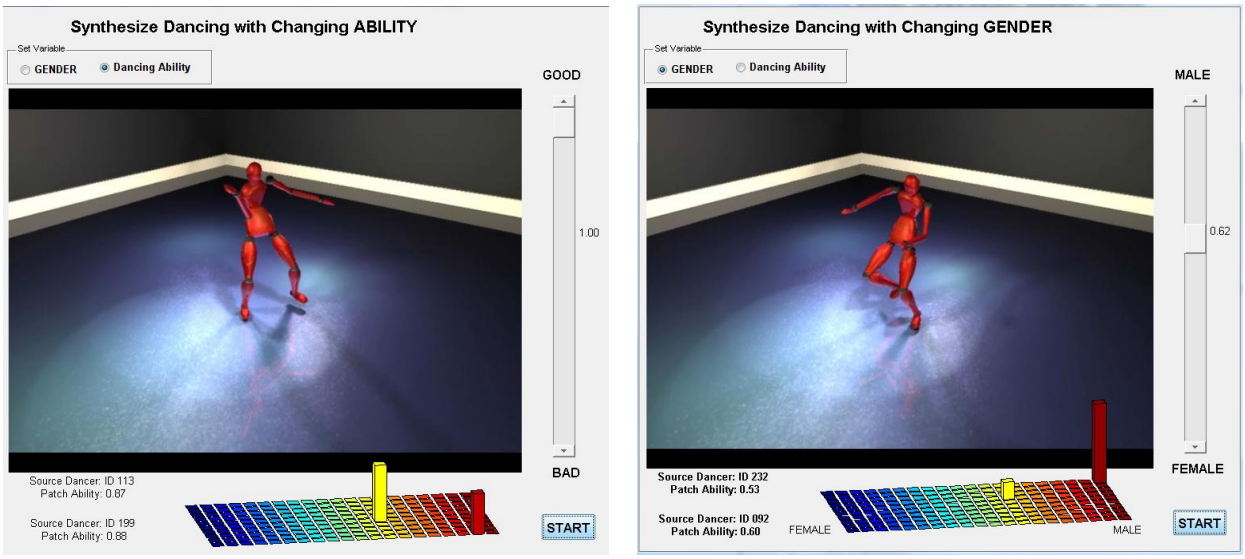


Figure 7: The user interface where the dance ability and gender of the synthesized dancer can be controlled and visualized. Each colored-square corresponds to a dancer with ranked (color-coded) label for dance quality or for dancer gender. The synthesized dance transits across smoothened dance texels/patches from different dancers (see the two color bars in transition coming out for the plane.

ture synthesis. The analogy between creating new, targeted dance animations from existing mocap data and image-based texture synthesis on 2D textures is motivated by the similarity of a common goal: to generate new (synthesized) texture samples that are *visually* and *statistically similar* to the given texture's regularity [11, 19, 22]. Dance texture/texel plays the similar *dual function* of representing spatial transformations (18-body-joint motion space) as well as a 2D color-image such like the *geometric deformation-field texture* does in the near-regular texture synthesis algorithm [19]. We map back and forth between a 2D dance texture and a motion space, guided by the intended/targeted semantic ratings.

Different from general texture synthesis however, the regularity of a music texture guides us to segment a dance texture into unit-length dance texels. The basic idea for our dance texture synthesis algorithm is: given a desired level of dance ability (gender), our algorithm generates a piece of dance texture that meets the goal by choosing from the candidate dance texels with the highest *visual similarity* as well as the closest gender or dance quality compatibility.

Dance Texture Synthesis Method

We have developed a novel similarity function for implementing a dance texture synthesis algorithm, adapted from image quilting [11] and near-regular texture synthesis [19, 22]. Here we define a *texture patch* as a pair of consecutive texels for the use of feathering during the synthesis process. The two-texel-length patches are used to achieve feathering for smooth transitions between dance patches overlapping on one texel. The similarity function for the best matched patches requires: 1) compatibility between consecutive patches and 2) class (ability or gender) and level (high or low) compatibility of each patch. The compatibility of the consecutive patches determines the smoothness of the dance; if tolerance in pixel-wise overlapping texels difference is too high (e.g. > 0.5 degree), the resultant dance motion may contain abrupt rotations. Hence, we first randomly pick and filter dance patches for compatibility to yield smooth body motion transitions. Second, as the classification of each patch determines the perceived rating of the entire dance, we filter out patches with undesired rankings. In addition, higher weights are placed on the more perception discriminative body joint features (Figure 5), leading to a body joint-sensitive similarity metric. Figures 7 and 8 show the dance texture synthesis process and the interface where the dance ability and dancer's gender are controlled independently. The texture synthesis process has three steps: 1) patch-finding for most similar texel pairs, 2) patch pool screening for a specified targeted rating, and 3) feathering adjacent patches for smooth dance movement transition.

Synthesize Dance To New Music

Let the basic period of origin music be T_0; to synthesize dancing with music having a different basic period T we resample (warp) the dancing texture with ratio T/T_0 to make the dancing faster or slower to go with the new music. The same alignment is done as described in the pre-processing stage.

Sample movies of synthesized dancers can be found here: http://vision.cse.psu.edu/research/MTurkDancing/index.shtml

3.4 Human Perception Validation

We validate the results of our dance synthesis with crowdsourcing. Mechanical Turk (MT) provides a general platform for deploying units of informational labor (called human intelligence tasks, or HITs) to users on the Internet who are paid for their efforts. Our user interface, built in Flash, functions in the same way for the dance quality and dancer gender assessments (Figure 9). At the beginning, the user is shown a sample video that contains thumbnail versions of four dances. The dance quality of these dances cover the extreme ends of our 1-5 range. The user is required to watch the sample video for 20 seconds and can then begin the labeling task. For each assessment, a single video is shown. The user can click the quality or gender radio buttons at any time and change the answer any number of times, but the 'Next' button is disabled until 20 seconds have elapsed. The user repeats this for 10-12 assessments before the HIT is over. The order of the sample videos is randomized per user. For the gender assessment task the order of the videos is randomized and the order of male/female buttons is also randomized at the start of each HIT.

To ensure rating consistency, we have each user perform assessments twice, though the user is not informed of this fact. This is similar to an approach used previously [9] in which all shape perception HITs were presented twice to each MT user as means of assessing consistency. In our case, we ask for 5 or 6 unique assessments, and each appears twice. Numerical consistency of the user is computed as the average of the squared differences between each pair of responses. The number lets us assess user consistency on a relative scale per task. We can then remove all data from users who have highly inconsistent scores.

A user could complete the task in a trivially consistent way by always clicking the same response. Therefore, we also look at the variance of a user's assessment scores. If the variance is too low, the user is either not attempting the task in earnest, or is failing to see significant differences between the input videos.

A final metric we can use for assessing the work quality is the 'click time.' Because we built the assessment interface, we are free to track numerous details of the user interaction, including timing information and tentative, non-final responses. If the user clicks a particular answer, then changes the answer, this is recorded and sent to our server, including the amount of time that has passed with each click. We want the user to form his/her opinion based on a reasonable length of video, not just a couple seconds. So we can throw out any work by users who click their final assessment too quickly. We do not penalize users for making multiple assessments or for making a non-final assessment too quickly.

To turn the aggregated MT results into an assessment of our ability to synthesize videos, we look at the correlation (R) between our targeted scores and MT assessed scores. We perform thresholding as described above to exclude disqualified HIT results, and then compute the correlation on the remaining values with a null-hypothesis that there is no correlation between the intended synthesized dancer quality/gender and human ratings. The numerical results shown in Table 2 suggest strong agreement between our synthesized dancers ability/gender and human raters; especially for dance quality assessment where the correlation score is strongly linear (R-value = 0.7251) and statistically significant (p-value is very low). Dancer's gender assessment from the MTs shows significant positive correlation with the computer targeted dancer genders, but the linear relationship is only moderate (R-value = 0.451).

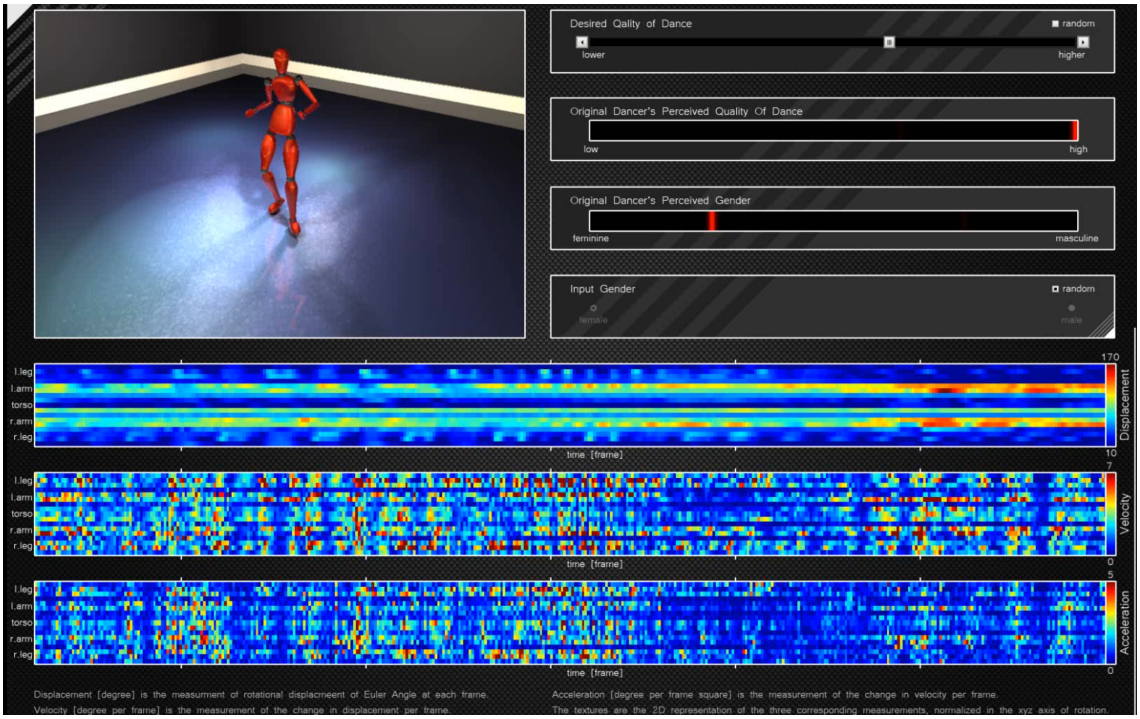


Figure 8: Dance-synthesizer at work: the bottom three rows demonstrate the three dynamic dance textures (Displacement, Velocity and Acceleration); top right displays the user desired dance quality; the two sliders below show the original data point (dancer) information of both perceived dance quality and perceived dancer's gender. Movies of synthesized dancers can be found here: http://vision.cse.psu.edu/research/MTurkDancing/index.shtml

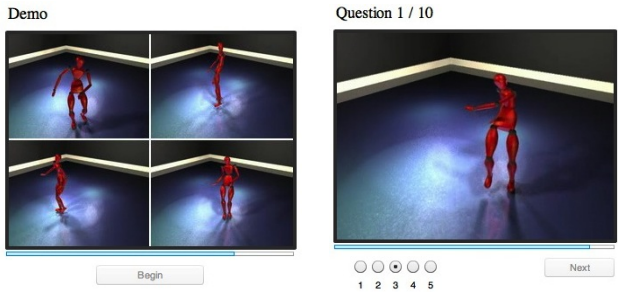


Figure 9: The interface for Mechanical Turk

Table 2: Summary of the Mechanical Turk Assessments of Synthesized Dancers

	Dance Quality	Dancer Gender
Total HITs	302	260
Unique workers (effective size)	232	189
Accepted HITs	244	200
R-value	**0.7251**	**0.451**
P-value	3e-216	0

4. SUMMARY AND DISCUSSION

We have proposed an effective framework for music and motion capture dance analysis, synthesis and evaluation. We have built a dance-synthesizer (Figure 8) and we have validated this framework on a mocap dataset containing 172 dancers, an asychronized piece of music, 25 initial human raters and more than 400 independent online raters.

Quantitative results from Table 2 provide a justification of our approach and its outcome. Starting with human input, analyzing the mocap data in a weakly supervised manner, and synthesizing the dancers based on learned discriminative features, we have produced promising output (synthesized dancers with certain targeted ability/gender scoring) that are confirmed by the human raters at a statistically significant level.

4.1 Motivated Raters

From a large amount of enthusiastic feedback of MTs (samples are shown in Table 3), it seems obvious that the raters are highly motivated: they personalize the computer (synthesized) animated dancers and treat the evaluation as a form of entertainment. These positive reactions are encouraging signs that *dancing with Turks* has a high potential to become a human-computer interactive game. With the support of the advanced technology, we expect our work lead to a rewarding experience for both the dancers and the actively participating audience (raters).

4.2 Dancer Ability versus Gender Perception

To our surprise, human perception of gender from dance animation (motion alone) of real dancers turns out to be rather challenging. From the 25 initial raters, the human

Table 3: Sample Comments from Anonymous Mechanical Turks (Gender, Age-range, Country, Self-rating, Comment)

F,31-40,CA,4,	"One of the most interesting HIT's I've done yet. You have made me very curious as to your study/project."
M,25-30,IN,3,	"Its bit difficult to find out the gender. "
F,18-24,US,1,	"The videos looked super cool!"
M,25-30,IN,4,	"Awesome job TNT"
F,31-40,US,1,	"I think the males tend to dance with pelvis forward"
M,25-30,US,3,	"Most of the dances seemed to have a salsa tone to them."
F,18-24,US,4,	"This hit was AWESOME!!!!"
M,31-40,IN,4,	"WANT MORE HITS LIKE THIS"
M,18-24,AU,2,	"Love the last dance."
F,31-40,IN,3,	"It was a nice entertainment."
M,18-24,IN,4,	"great animated dance"
M,31-40,BE,1,	"This was a nice HIT to do. What will you use it for exactly?"
M,18-24,PH,3,	**"the hands determine my choice of the dancers gender."**
F,18-24,US,1,	"The music used is addictive!"
M,41-50,US,5,	"professional theatrical dancer. None earned a 5 for me. Criteria Used: **level of energy exhibited, range of motion, variety in patterns**"
M,18-24,CA,4,	"good movements , to me the **best ones were the ones where it starts slow and requires a lot of foot work** with some body movements as it goes with the music"
F,41-50,US,2,	"amusing"
M,31-40,CA,3,	"great job! Lots of fun !"
F,31140,CA,4,	"Great HIT! Don't know if it helps, but I find usually women are more inclined to make larger movements, and cover more floor space than men when dancing. "
M,25-30,IN,3,	"Its bit difficult to find out the gender. However,i have done my best.Tnx"
M,51-60,US,2,	"Wow, very interesting!!!"
F,18-24,US,1,	"Fun to watch!!"
M,18-24,PH,4,	"Awesome!These is great!Love it very much.Break hit!"
M,41-50,US,1,	"I thought the animations were awesome and I actually was learning moves by watching"

gender classification rate is 64.78% ± 7.6% with approximately equal performance on both genders. For computers, discriminating and synthesizing gender specific subjects is also less agreed up by MTs than controlling the levels of dance quality (Table 2). This challenge may arise from the nature of gender as a binary variable; a 0.5 rating value in ability means the dancer is an *average* dancer, while a 0.5 value in gender means 'either' male or female, therefore the variance is potentially much higher. In the context of relatively easy recognition of gender through human gaits [14], this issue of gender recognition in free-form dancing is of high social and scientific interest.

4.3 Future Work

So far, we have explored the gender and dance ability of a dancer separately. Future work to test our methodology will include both gender and dance ability jointly. Finding the dance texture patch that matches both ratings the best within the training data pool allows the resultant motion to have more than one controllable characteristic.

Furthermore, although we closed the loop in this work in one cycle, our learning-evaluation cycle can continue evolving. Synthesized results can further be reclassified by human raters (Section 3.4) and fed back into our analysis/learning system. This will allow the machine to learn and adjust itself accordingly over time, as if in a back-and-forth dance with the *Turks*.

5. ACKNOWLEDGMENTS

I-Kao Chiang (first author) worked on this project for his BS honors thesis while at PSU. Brian VanLeeuwen contributed to Figure 3(A). We thank the authors of [6] for sharing their motion capture data. This work is supported in part by NSF grants IIS-1248076 and IIS-1144938.

6. REFERENCES

[1] O. Arikan and D. Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics*, 21(3):483–490, 2002.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995. ISBN:0198538499.

[3] M. Brand and A. Hertzmann. Style machines. In *Proc. Conf. on Comp. Graph. and Int. Techn.*, pages 183–192, 2000.

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH 97*, pages 353–360, August 1997.

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8–15. Citeseer, 1998.

[6] W. M. Brown, L. Cronk, K. Grochow, A. Jacobson, K. Liu, Z. Popović, and R. Trivers. Dance reveals symmetry especially in young men. *Nature*, 438, 2006. Retracted December 2013.

[7] A. Bruderlin and L. Williams. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, page 104. ACM, 1995.

[8] C. Cedras and M. Shah. A survey of motion analysis from moving light displays. In *1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94.*, pages 214–221, 1994.

[9] F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh. How well do line drawings depict shape? In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, volume 28, Aug. 2009.

[10] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.

[11] A. Efros and W. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, pages 35–42, 2001.

[12] K. Grochow, S. Martin, A. Hertzmann, and Z. Popović. Style-based inverse kinematics. *ACM Transactions on Graphics (TOG)*, 23(3):522–531, 2004.

[13] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In *Proceedings of the 32nd annual conference on computer graphics and interactive techniques (SIGGRAPH 2005)*, pages 1082–1089. ACM Press, 2005.

[14] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perceiving events and objects*, 3, 1973.

[15] L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. *ACM Transactions on Graphics (TOG)*, 21(3):473–482, 2002.

[16] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

[17] J. Lee, J. Chai, P. Reitsma, J. Hodgins, and N. Pollard. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics*, 21(3):491–500, 2002.

[18] Y. Li, T. Wang, and H. Shum. Motion texture: a two-level statistical model for character motion synthesis. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 465–472. ACM, 2002.

[19] Y. Liu, W. Lin, and J. Hays. Near-regular texture analysis and manipulation. *ACM Transactions on Graphics (SIGGRAPH)*, 23(3):368–376, August 2004.

[20] Y. Liu, K. Schmidt, J. Cohn, and S. Mitra. Facial asymmetry quantification for expression invariant human identification. *Computer Vision and Image Understanding Journal*, 91(1/2):138–159, Special issue on face recognition, Martinez, Yang and Kriegman (Eds.). July/August 2003.

[21] Y. Liu, L. Teverovskiy, O. Carmichael, R. Kikinis, M. Shenton, C. Carter, V. Stenger, S. Davis, H. Aizenstein, J. Becker, O. Lopez, and C. Meltzer. Discriminative mr image feature analysis for automatic schizophrenia and alzheimer's disease classification. In *7th International Conference on Medical Imaging Computing and Comptuer Assisted Intervention (MICCAI 2004)*, pages 378–385. Springer, October 2004.

[22] Y. Liu, Y. Tsin, and W. Lin. The promise and perils of near-regular texture. *International Journal of Computer Vision*, 62(1-2):145,159, April 2005.

[23] E. Moulines, P. Emerard, D. Larreur, J. L. S. Milon, L. L. Faucheur, F. Marty, F. Charpentier, and C. Sorin. A real-time french text-to-speech system generating high-quality synthetic speech. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1990.

[24] R. Murray, Z. Li, and S. Sastry. *A mathematical introduction to robotic manipulation*. CRC, 1994.

[25] W. Pan and L. Torresani. Unsupervised hierarchical modeling of locomotion styles. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 785–792. ACM, 2009.

[26] K. Pullen and C. Bregler. Motion capture assisted animation: Texturing and synthesis. *ACM Transactions on Graphics (TOG)*, 21(3):508, 2002.

[27] T. Qin, X.-D. Zhang, T.-Y. Liu, D.-S. Wang, W.-Y. Ma, and H.-J. Zhang. An active feedback framework for image retrieval. *Pattern Recogn. Lett.*, 29(5):637–646, Apr. 2008.

[28] C. Rose, M. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998.

[29] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa. Video textures. In *Proceedings of ACM SIGGRAPH 2000*, pages 489–498, July 2000.

[30] S. Sumi. Upside-down presentation of the Johansson moving light-spot pattern. *Perception*, 13(3):283–286, 1984.

[31] G. W. Taylor and G. E. Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proc. Int. Conf. on Mach. Learn.*, pages 1025–1032, 2009.

[32] L. Torresani, P. Hackney, and C. Bregler. Learning motion style synthesis from perceptual observations. In *Adv. in Neural Inf. Proc. Sys.*, pages 1393–1400, 2007.

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

[34] R. Urtasun, P. Glardon, R. Boulic, D. Thalmann, and P. Fua. Style-based Motion Synthesis. *Computer Graphics Forum*, 23(4):1–14, 2004.

[35] J. Wang, D. Fleet, and A. Hertzmann. Multifactor gaussian process models for style-content separation. In *Proc. Int. Conf. on Mach. Learn.*, pages 975–982, 2007.

[36] A. Witkin and Z. Popovic. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 105–108. ACM, 1995.

[37] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.