Viewer Preference of Aesthetic Pose Quality in a Gesture Corpus

Topher Maraffi, Sascha Ishikawa
University of California, Santa Cruz
Santa Cruz, CA 95064
[topherm, stishika]@soe.ucsc.edu

ABSTRACT
This paper describes a computational aesthetics approach to inferring the quality of poses in a gesture corpus. We describe experimental design, feature implementation, and a pilot study. Motivation is taken from computational cinematography studies for virtual camera control. We apply similar methods to virtual character control, with the goal of enhancing avatar and NPC acting in cinematic videogames. Design includes Poserama, a gestural game designed to facilitate data collection to study character pose preferences by media consumers. Aesthetic features are based on arts and animation theory, and our algorithms score pose features along three dimensions: balance, asymmetry, and readability (BAR). The pilot study was run on a pre-processed corpus from motion capture data, with some encouraging results. This work is a first step in a computational performatology approach to skilled acting affordances in videogames, by learning the aesthetic features of affective pose representation in the arts.

1. INTRODUCTION
Popular videogames that feature cinematic style stories and characters are creating new opportunities for performative play. Games that vary in genre, narrative, and mechanics often display the common operational logic of fictive characters being controlled by either human players or AI agents. In non-interactive visual media, the role of playing characters has always been the prerogative of skilled artists because audiences rely on the aesthetic quality of gesture to follow the flow of meaning in moving pictures. Performing characters, whether through live acting or more indirect animation methods, leverages aesthetics from the Performing Arts domain that were designed to affect viewers. For interactive videogames, players are both performers of avatars and spectators of non-player characters (NPCs), so how characters are performed is becoming central to the aesthetic experience of this type of interactive media.

The availability of full-body gestural controllers using natural user interface (NUI) technology is moving role play in videogames towards physical acting. Since high quality character performance is related to artistic skill, it becomes increasingly important to better understand aesthetic conventions employed by artists to make embodied gesture appealing to spectators. We are motivated by three intuitive assumptions:
• Professionals trained in artistic conventions produce higher quality popular art than untrained amateurs, and indicate the skill of the artist through aesthetic features.
• Exposure to professionally produced media has created aesthetic literacy in the general public, so that popular preferences reflect artistic conventions, and spectators can distinguish between pro and amateur quality in popular art.
• In videogames, animated avatars and NPCs can perform at a higher quality level by training players and agents to incorporate Performing Arts conventions, making interactive characters display features of skilled acting.

Although we are applying these assumptions to performed gesture in media, we are also drawing from research in photography and cinematography. Professional photographers learn composition rules that indicate pro quality to the general public, and it is assumed that amateurs can improve their photos by learning to apply such rules. Similar studies on visual composition quality for camera control implemented mathematical approximations of photography rules to execute virtual camera directives that achieve optimal scoring [Swanson et al. 2012]. In camera aesthetic meters attempt to improve amateur photos through real-time feedback for inferred quality [Joshi et al. 2011]. Our performatology approach is similar in that we want to identify features that contribute towards the compositional quality of a gesture, and assess their relative importance in predicting perceived quality of character acting in games. Gamers are media savvy, so they are literate in cinematic acting conventions, and should be able to distinguish differences in gesture quality. Beyond evaluating the aesthetic experience of character gesture in media, applying a Performing Arts model to role play in games will facilitate computational tools that enhance player expression through creative acting affordances.

Aesthetic gesture evaluation is an important aspect of cinematic videogames because character interaction often precedes communicating narrative in visual storytelling. Story content builds out of interaction, with players engaging characters even before narrative information is conveyed. As play in games becomes more performative, players gain more opportunities to improvise character content not entirely scripted by the game designer. A large part of live acting in theater and film is developing the improvisational skill to quickly generate gestural content not specified in the script. Actor motivation, or what a character should do, is provided by the script author and director. But the delivery, or how the character acts, is left to the performer to create in the moment. Skilled actors use interesting gesture to generate personality for a stereotypical character type that makes them interesting to watch. Arts literature contends that gestural conventions are designed to catch and hold spectator attention through visual interest [Barba and Savarese 2005].
contrast, everyday gestures are intended to efficiently complete a task without regard to spectator affect.

At a base level, continuous gesture in media breaks down to a series of captured or rendered poses. This is true in film and animation, where the frame or keyframe is the atomic element of visual information. This is also present in choreography for live performance, where poses are planned as moments of relative stillness compared to transition gestures, and have greater semantics importance. A professional performer like Michael Jackson displayed very clear iconic poses that are often imitated to convey his recognizable stage personality. Gestural practices in the Arts train performers to have greater awareness and control over how they present their body to an audience. Physical exercises teach them to isolate the movement of individual body parts in order to convey a series of interesting shapes. Disney animators, who studied filmed actors in the early part of the 20th century, derived principles for animated characters that are directly related to pose composition [Thomas and Ollie 1981]. Lead artists would apply these principles to keyframe drawings to make the gesture more readable, appealing, and interesting. Therefore, before studying the aesthetic features of continuous gesture, we focus on understanding how performers control their body to strike an appealing pose.

For understanding pose composition, we need to develop models that compute the visual quality of a pose according to embodied acting conventions, and ideally also for individual player preferences. Our Poserama game is designed to generate a corpus of poses that are scored and annotated for key features from domain literature. Similar to photo composition ranking, pose composition can be rated for visual features related to balance, thirds alignment, and symmetry. The goal is to understand what makes skilled performers interesting to watch, so that interactive characters can be designed to optimize these features in their poses. To raise the aesthetic quality of embodied characters in games, we require pose generation algorithms that model aesthetic principles from the Performing Arts, but so far no such tools exist. We attain our initial feature set from character principles in acting and animation theory related to pose composition, which include gesture practices that feature dynamic balance, asymmetrical lines, and silhouette readability. Our features are used to score poses in a gesture corpus, initially from pre-recorded mocap data, and later from streaming 3D data captured by a Kinect camera during game play.

2. RELATED WORK

With the exception of a couple of cases [Bates et al. 1993, Magerko et al. 2010, Seif El-nasr 2004], computational work on how artists represent embodied character behavior is sparse. Animation approaches to derive gesture style have treated artistic performance as pre-processed data to be manipulated and blended [Oshita 2008], rather than artistic behavior rules to be learned. Intelligent virtual agent approaches to procedurally generated behavior tend to model natural gesture of everyday people [Kipp et al. 2007] rather than model the artificial techniques of skilled performers. Instead, our approach in this paper more closely relates to work in computational inference of aesthetics and player preference modeling for intelligent virtual camera control. This may seem like an unusual match, but the goals of cinematography and performance are both aimed at optimizing visual aesthetics for spectator experience. Actors predict how the lines of their body will read when framed by the proscenium in theater, or when captured by the camera on film, and compose their gesture for maximum dramatic affect on the spectator. Animators create keyframe poses that incorporate aesthetic principles to make them readable and appealing. Intelligent virtual character control modeled on how artists present the body will need to calculate the most aesthetically interesting poses in relation to the virtual camera viewpoint. We hypothesize that similar design rules for generating visual interest inform composition procedures in both cinematography and performance.

An image processing tutorial on the emerging research area of computational inference of aesthetics and emotions in images was done by [Joshi et al. 2011]. Image quality assessment and semantics inference studies draw from aesthetics in the visual arts, especially photography and painting, as well as from psychology work on the states of mind involved in the aesthetic experience. Key challenges include an “aesthetics gap” that stems from the inherent semantics gap between low-level computable visual features and high-level subjective human semantics. They point out that the experience of visual art is related to viewer experience, with skilled artists having a different perception of aesthetics than less knowledgeable spectators. Cross-cultural studies have shown that the general public has less appreciation for conceptual or abstract art, placing more emphasis on visual aesthetics than the underlying ideas. They recommend that computational aesthetics statistically model the observed aesthetics of an artwork under a constrained population of experts and general viewers. Systems that reliably infer responses to popular art would, for at least a section of the population, assist in decision making tasks related to media. This includes searching and sorting through large collections of images according to a quality or appeal measure. They analyze several Web-based photo resources that crowd source such measures, including Flickr's that has a dynamic interestingness feature, and sites that implement peer-based quality ratings like Photo.net and DPChallenge.

Joshi et al show that computational frameworks for this area of study have been influenced by psychology methods and visual art theory. Art theorists hypothesize that there are multiple levels of aesthetic perception, starting with surface resemblance to objects, followed by a deeper level of cognitive interpretation that depends on the cultural environment and habitual semantic rules. Some theorists argue the importance of artistic intention, and for standards of correctness in pictorial representation to evoke an aesthetic experience. In a brief history of photography, they show that Pictorialist photographers in the 19th century borrowed common techniques from painting to develop photography as an art form. Formal composition rules, like the Rule of Thirds, have been displayed in the popular arts of painting, photography, and cinema. Such rules have been implemented as quality inferring features that computational studies have used to distinguish professional from amateur photography. Another common artistic technique of landscape photographers pointed out by Joshi et al is to use film that exaggerates color intensity, such as Fuji Velvia, rather than film that produces true to life or natural colors. Studies show that color intensity, contrast, and bluriness are also features that infer professional quality in photos.
The Gestalt psychology concept of goodness configuration, where perception is organized according to properties like symmetry and simplicity, has also influenced feature representation in aesthetic inference studies. Other related concepts include Physiognomics and Neuroaesthetics, which investigate psychophysiological mechanisms of aesthetic perception. These include analysis of eye gaze for fixation and movement patterns, and other neurological studies of the perceiver’s processing dynamics. Processing fluency theory suggests that the more fluently a perceiver can process an image, the more positive their aesthetic experience will be. Core problems for computational studies are predicting aesthetic and emotional responses for cliques in the general population, and understanding individual preferences that make some images more appealing than others. Joshi et al noted that a regression framework for scoring aesthetic and emotion prediction is highly challenging due to noise in user ratings. Instead, a high versus low quality binary classification makes the problem more tractable, which can be evaluated by standard accuracy measures. For predicting appeal, they recommend that personal and situational preferences should be taken into account, as well as social networks. Examples of network reinforced appeal can be seen in Facebook “likes” and Flickr’s “interestingness” attribute.

The dominant machine learning strategy for aesthetic inference has been supervised support vector machine (SVM) classification, with classification and regression trees (CART) for emotion prediction. Joshi et al consider aesthetic inference as more abstract than emotion inference, making it more difficult to reliably classify into a finite number of classes. However, the bulk of aesthetic and emotion inference approaches have used classification, with a two-class aesthetic prediction for “high” versus “low” quality being the most tractable approach, and a similar “interesting” versus “boring” scheme for emotions. Unsupervised learning has also been done using K-means clustering for visual vocabulary generation, graph-based region segmentation, and image clustering to form topical groups. Bayesian network classifiers have also been used for generative learning, and to classify facial expressions to infer emotions. They list several real-world applications of visual learning metrics for photographic systems, including real-time camera filters that feature an aesthetics meter.

A recent example of a top-down binary classification approach to photographic ranking was done by [Yeh et al. 2010]. They implemented a personalized ranking system that featured algorithms based on aesthetic rules of photography for composition and color distribution. Crowd sourced ratings data from the DPChallenge web site was sorted into “preferred” and “non-preferred” classes, and then ListNet was used for feature extraction and training their binary classifier. Their implementation used a “late fusion” method to produce an aesthetic score as a linear combination of optimal weightings for their top features, such as simplicity and contrast, which then produced a ranked list with a reported 93% accuracy. Users were presented with an initial ranking of photos in two types of interfaces: one where they manipulated feature sliders to re-rank the photos, and another where they clicked on preferred example photos for automatic re-ranking. Included in the interfaces were more subjective personal features like portrait, color, and aspect ratio. Although these features did not relate to quality conventions, they were designed to appeal to the personal taste of the amateur photographer. Users reportedly preferred the more visual example-based interface to the slider interface by two to one.

Other recent work that specifically focuses on evaluating quality in amateur portrait photos was done by [Khan and Vogel 2012]. Their top-down approach reduced the classification space to only 7 composition and color features that specifically target portraiture. They make the point that most statistical methods take a bottom-up approach by calculating a large amount of global features across a general space of images to correlate a subset of relevant features to visual aesthetics. In a template-based method they combine composition principles like the rule of thirds and golden mean into a two dimensional lookup table, and claim better classification results with 1/9th the size of the traditional feature set. The human photo data set with ratings from Mechanical Turk was their training data. Using the median of all ratings for ground truth, they labeled portraits for a binary classification of “high” and “low” quality using five Weka algorithms in a 10-fold cross validation process. The top performing features were all composition related, with rule of thirds being consistent with other more general feature sets.

The closest related work to our project, however, is a mixed approach that uses composition rules to help generate and annotate images with aesthetic features, and then employs statistical methods to correlate the features with viewer preferences [Swanson et al. 2012]. They used gameplay scoring to help generate landscape images that were annotated according to aesthetic rules from photography texts, and then implemented crowd sourcing on the resulting corpus to correlate viewer preferences. Color coded badges in the game functioned as an aesthetic meter, encouraging gameplay towards the goal of taking pictures that rated highly for the three main features of thirds, balance, and symmetry. Generating images in the target domain reduced the feature set in a way that would be difficult using a corpus of photos from image repositories on the Web. This gave them additional control over feature dimensions through design abstraction, such as removing color or minimizing cultural references in the game’s representation of landscape photography. User studies through Mechanical Turk were then classified for quality using pairwise comparisons of four-alternative forced choice (4AFC) preference ratings to learn individual as well as general preferences. Their multi-class SVM algorithm got the best accuracy by reducing the choices to a binary classification task where a clear choice indicated one image was “good” and the other was “bad”. They reported relatively accurate predictions for an arbitrary user, and even better accuracy for individual users with less training examples.

3. DATA GENERATION IN POSERAMA

After an initial pilot study that used a pre-processed gesture corpus, data generation in Poserama is performed live as part of a Kinect game on fashion photography. Representation in the game is minimalistic in order to target the mechanic of aesthetic figure composition, and reduce natural, cultural and gender references. Captured poses are classified and annotated according to composition quality metrics informed by arts and animation theory. Player motivation is in the form of color-coded badges that function as aesthetic meters, scoring for the target features of balance, asymmetry, and readability. Additional performance
feedback is through sounds and animation during the pose capture phase of the game.

3.1 Game Design
Poserama is a Kinect game where the player performs the role of a fashion model by striking poses for the camera. We implemented a minimalistic design approach similar to the Panorama game [Swanson 2012] by removing elements of color, texture, and cultural specificity from the scrolling background. The androgynous 3D figure is also rendered in grayscale with minimal features. The figure automatically animates towards the left on a walk cycle as a generic urban city street scrolls from left to right. The walking figure has the performance attitude of a fashion model on the runway, with an accompanying upbeat soundtrack with samples of ‘on the catwalk’. Photo opportunities are delineated by a darkened frame surrounding objects in the environment. Upon entering the photo op, the frame expands and lightens while control over the figure is given to the player, and a 10 second countdown for the camera timer starts. Game mechanics require the player strike an interesting pose within the frame using the Kinect camera. Flash bulbs go off, with audio encouragement for a “good” pose from the photographer and fans, while three successive pictures are taken. The best scored of the three captured photos are displayed in a film strip across the top and bottom of the screen, which can be examined by the player at any time by clicking on it.

Each pose is automatically rated for features of balance, asymmetry, and readability, or BAR, with player feedback on the quality of their pose indicated by three colored badges in the upper right corner of the game. The score in the main window is cumulative, with red indicating a low rating and gold indicating a high rating. Players are not told what the individual BAR badges mean, so they must infer how to compose their bodies to produce a high score. After the third image is taken the figure continues walking; only now the walk cycle reflects their game score. A high score produces a confident model walk with upbeat music, while a low score produces a hesitant walk with sluggish music.

Level 1 of the game continues through 10 photo opportunities, capturing a total of 30 poses, which are ranked and annotated according to their feature scores. 10 poses from each of the low and high quality classes are then randomly selected to add to our corpus. In addition to feature annotations, in the setup part of the game we ask them to self rate, from amateur to professional, their experience level in modeling, acting, and dance. This allows us to compare global scores between amateurs and professionals to verify the intuition that skilled performers will consistently score higher according to our metrics, and normalize rankings in subsequent user studies.

4. BAR: SCORING FOR GESTURAL FEATURES
Many of the composition rules used in photography are also utilized in the related domains of cinematography and graphic design. For instance, the rule of thirds can be generalized as a structural feature of any framed composition, including ones that feature figures. Other rules used to infer aesthetics in images, such as balance and symmetry, have comparable principles in embodied art practices like acting and dance. We hypothesize that composition rules in graphic and figurative arts have similarities because they have the same goal of catching and holding the viewer’s attention through visual interest. Good composition through the arrangement of shapes and lines stimulates eye movement through the framed image space, which reflects visual processing theories from psychology [Reber et al. 2004, Peters 2007]. Pose composition rules are based on general conventions in the performing arts that make a performer interesting to watch even before narrative information is conveyed, and is associated in theatre with stage presence. In studying the aesthetic quality of pre-narrative poses, we hope to better understand how a posed body can engage viewer attention in order to establish a communication channel for character and narrative content.

Using a top-down approach to feature selection, we used art and animation theory to reduce our pose features to three metrics: Balance, Asymmetry, and Readability (BAR). Since we are interested in scoring the aesthetic quality of posed figures, we chose expressive features that are displayed by a variety of professional performers, including actors, dancers, models, and even animated characters. Our assumption is that skilled players trained in gestural performance conventions will consistently score higher for our metrics than unskilled players, and therefore will produce poses that rank higher for quality by the general public. This implies that consumers of popular media are literate in these conventions, and can learn to be better avatar performers through real-time pose feedback that functions as an aesthetic quality meter.

4.1 Balance
In performing arts theory [Barba 2005], skilled performers train at pushing the limits of their balance. Poses that are evenly weighted are considered static and boring, while poses on the edge of imbalance, without falling, are considered dynamic and interesting. Dynamic poses often shift the hips and shoulders in opposite directions, achieving an S shape, which is considered aesthetically pleasing. Physically, if the upper portion of the body bends horizontally, the lower portion must counter in the opposite direction to not fall over. This pose principle has similarities to photo composition rules, in that body parts are moved off the center line to power points in thirds regions, and create lines reminiscent of the golden triangle, as used in other works on aesthetic inference [Khan 2012, Swanson 2012].

![Figure 1. Design for the torso Balance pose feature based on arts and animation theory.](image-url)
The balance feature seeks to capture an intuitive notion of the pose difficulty by measuring the horizontal distance between the upper and lower parts of the torso. This was implemented so that poses that push the boundary of balance, without falling, will have more aesthetic value. It should be noted that from a physics and mechanics perspective, our particular use of the term “balance” relates to the body’s center of mass rather than the body’s dynamic balance, or equilibrium. As a result, we make the assumption that the performer’s body is always balanced in this sense – that the forces acting on the body’s center of mass cancel each other. Otherwise the performer will fall.

In order to account for the figure’s use of horizontal space, we consider the distribution of joints along a central vertical axis that bisects the horizontal span of the pose. This measure, which is reminiscent of the rule of thirds in much as the pose is evaluated with respect to its “spatial composition,” was implemented by calculating the average horizontal distance of selected joints to the central axis.

4.2 Asymmetry

Another general principle displayed by skilled performers is asymmetry in posing their limbs. An asymmetrically posed body has variations in limb rotations across the Y axis, so that one limb may be straight while the other is bent. These create expressive lines that move the eye of the viewer across form of the figure. In performing arts theory [Barba 2005], symmetrical poses are considered static and boring, while asymmetrical poses express character. Symmetrical poses may be easier to read because the same content is reflected on each side of the body, but asymmetrical poses are information rich and display a higher degree of physical control. Recent psychology studies on symmetry and facial beauty have shown that functional asymmetries in the natural faces of models scored higher for attractiveness than digitally mirrored faces [Zaidel 2005].

Figure 2. Illustration of a balanced (left) and unbalanced (right) pose. The central axis is shown as a blue vertical line. The average value of horizontal distances between selected joints (filled in red) and the central axis are used in the balance feature.

\[
\text{Asymmetry} = \sum_{j=1}^{N_j} \left( \sqrt{(\theta_{L,j}^{(j)} - \theta_{R,j}^{(j)})^2 + (\theta_{L,j}^{(j)} - \theta_{R,j}^{(j)})^2 + (\theta_{L,j}^{(j)} - \theta_{R,j}^{(j)})^2} \right)
\]

where \( N_j \) is the total number of joint pairs.

4.3 Readability

Readability is a gestural principle utilized by performers, and is an animation principle that relates to how expressive a pose is by how affective it is on a viewer [Thomas 1981]. Psychology studies have shown that how easily an image can be cognitively processed is related to aesthetic pleasure [Reber 2004]. Skilled animators often pay special attention to the silhouette of the character they are drawing, and actors are continuously aware of their body positioning in relation to the camera or viewer in order to have a pose that has maximum expressiveness. Depending on lighting quality, if the pose is unreadable, it cannot convey enough information to a meaningful affect. Therefore, limbs are moved away from the body by subtly exaggerating gesture.

Figure 3. Design for the limbs Asymmetry pose feature based on arts and animation theory.
Since pose readability depends on the location of the viewer, the distance is calculated on the projection of points onto the two-dimensional camera plane. Assuming that the highest readability occurs when an actor is facing the camera with limbs stretched outwards from the body, a reliable estimate of readability was implemented using the relative distances between points on the actor’s torso and limbs. This is motivated by the observation that such points provide an estimate of the actor’s orientation with respect to the camera, which is in turn is related to the amount of occlusion. For instance, a side view of a pose results in heavy occlusion of the torso.

The readability is obtained by computing 1) the distance between the left and right hips to the pelvis, and the distance between the left and right shoulders to the neck, and 2) the distance between the left and right hands and feet. The latter provides an estimate of limb occlusion whereas the former measures the orientation of the torso relative to the camera including chest and hip occlusion due to foreshortening effects.

5. PILOT STUDY
We conducted a pilot study to determine how user preferences could be modeled using the BAR features introduced in this paper. Poses were extracted from motion capture sequences and evaluated by users in terms of qualitative preferences. These were used, along with the BAR features for the same poses, to model the users’ preferences using machine learning methods.

5.1 The Motion Capture Data
A performer’s gesture was recorded using motion capture equipment. A total of 1820 frames were recorded, each containing the location of points on the body of the actor’s joints in three-dimensional space. Additionally, the rotation angles along each of the axes were recorded for each joint. The performance consists of a diverse set of motions and styles.

A total of 128 poses were sampled from equally spaced frames within the motion sequence. For each pose, participants were asked to answer “yes” or “no” in response to the question “is this pose interesting?” Furthermore, participants were encouraged to answer quickly through timing constraints in order to prompt intuitive judgments, rather than dwelling on contextual details.

6. RESULTS AND DISCUSSION
6.1 Qualitative Analysis of BAR Features
An initial qualitative assessment was made in order to determine the efficacy of using the BAR features to characterize human poses. BAR features for three different motion sequences, namely the “lateral stretch,” full-body rotation, and the “arm wave,” were inspected manually. These are compared to the baseline pose of simply standing while facing the camera. Figure 2 plots each BAR feature throughout the sequence of frames. The motion sequences examined are highlighted in red, green, and blue.

We begin by establishing a baseline for the BAR features during a simple standing pose: the actor stands in a relaxed position, with limited amount of motion, while facing the camera. This is captured by the frames 1400-1600 (no highlight) in Figure 2. As expected, readability and balance are high and asymmetry is low. Frames 190-335 (highlighted in red) correspond to movements resembling the “lateral stretch” where the actor stands facing the camera while the torso sways laterally. The actor sways three times, which is captured by the periodic motion in the balance and asymmetry plots.

Frames 675-850 (highlighted in green) capture the full-body rotation, which produces corresponding oscillations in the readability plot. As expected, high readability values coincide with frontal views whereas lower values coincide with side views where the joints are heavily occluded. The actor stands straight with elbows bent, resulting in a nearly constant balance.

Frames 1290-1400 (highlighted in blue) capture the “arm wave.” In this motion sequence the actor stands facing the camera and begins a wavelike motion in one arm that travels across to the other arm, then back again. The actor leans into the direction of the wave, which produces a small, but steady, change in balance that is clearly visible as a “spike” in the balance plot. Since the body is viewed from the front with little occlusion, the readability remains high and relatively constant. The asymmetry is also high because each arm is performing a different motion as the wave travels from one side to the other.
6.2 Feature Space Analysis
Figures 3-5 illustrate the distribution of 128 “good” (green) and “bad” (red) poses in the BAR feature space. The tight cluster of “bad” points suggests that poses considered uninteresting by users are relatively easy to discern. However, given the sparse distribution of “good” poses it is reasonable to expect difficulties in determining which poses are interesting to users. This is consistent with a common observation that bad performances are more readily noticed than good ones. We use machine learning to investigate this further.

Figure 6: Feature space distribution.

6.3 Preference Modeling

6.3.1 Top-Down Approach
We used the C4.5 decision tree implementation in the Weka machine learning library [Hall et al. 2009] to learn the participant’s preferences given binary ratings for each pose along with its BAR features. Using a 10-fold cross-validation procedure, the resulting decision tree was able to rate “unseen” poses with an accuracy of 79.7 percent. The results are summarized in the following confusion matrix in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Good Pose</th>
<th>Bad Pose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Pose</td>
<td>45</td>
<td>17</td>
</tr>
<tr>
<td>Bad Pose</td>
<td>9</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrix summarizing the binary classification of poses using a 10-fold cross validation procedure with a decision tree.

The number of “good” poses incorrectly classified as “bad” (17) is larger than the number of “bad” poses incorrectly classified as “good” (Table 1). This is consistent with our previous observation, where we visually inspected the distribution of poses in BAR feature space, that good poses are probably harder to predict.

6.3.2 Bottom-Up Approach
In order to evaluate a bottom-up approach to modeling pose preferences using BAR features we used principal components analysis (PCA) to automatically generate features using linear transformations of the large feature space that spans the raw motion capture data. Each motion capture frame is represented as a 120-dimensional vector (20 joints per frame, each containing three spatial coordinates and three rotational angles). Applying
PCA to the raw motion capture data and using machine learning algorithms with 10-fold cross validation consistently resulted in poorer results.

7. DISCUSSION & FUTURE WORK
This paper describes the motivation, design, and pilot study behind Poserama, a computational performatology approach to aesthetic pose quality inference and prediction for games. We showed the close connection to aesthetic studies in computational cinematography and virtual camera control, which informed our approach, to show that visual aesthetics generalizes across both domains. We detailed our top-down experimental design, using a small feature set derived from art and animation theory, to implement a pilot study with encouraging initial results, including a 79.7% combined BAR feature accuracy. In addition, our study briefly explored a bottom-up approach to characterizing poses by using PCA to reduce the motion capture dataset and generate meaningful features automatically, but was unable to accurately predict pose preferences.

Future work includes implementing the full Poserama Kinect game to collect data annotated with feature scores through gameplay. We also intend to implement a crowd sourced study to investigate the global preference of poses, using our game generated corpus and pairwise ranking. Asking participants to perform binary rankings of poses proved problematic because of a lack of context (“the pose is interesting relative to what?”). Future work would benefit from user studies that employ pairwise comparisons (e.g. 4-AFC) between poses. Lastly, we would like to implement studies that explicitly measure the aesthetic connection of cinematography to performance. In real-world media production, figure and camera composition often happen simultaneously through directorial feedback. Automatically syncing figure posing with camera framing, by summing aesthetic features, would theoretically produce the most appealing imagery for spectators.

8. REFERENCES