Metrics for Motion Editing

Paul S A Reitsma

May 15, 2001

Committee: Nancy Pollard, David Laidlaw, John Hughes

Advisor: Nancy Pollard

Abstract

Motion editing is increasingly being accepted as a means for expanding the range of motions available for animating characters with far less expenditure of time and effort than the original motion creation method. Adding flexibility and breadth to motion capture data is particularly valuable, provided the new motions are sufficiently "good." Computable metrics for determining whether a motion is good enough for a certain context would enable the automatic generation of edited motions to be done with much greater confidence, vastly increasing the range of practical applications. In addition, a sufficiently powerful suite of metrics could aid in examining the strengths and shortcomings of different methods of motion editing. This research opens the field of metric analysis of motion editing and presents some first, and encouraging, results.

1 Introduction

1.1 Motion Editing Overview

There are several widely-used methods for generating high-quality animated motions, with the most common being motion capture, procedural technologies, and simple keyframing. Each of these has certain strengths and limitations; motion capture, for example, is excellent at generating realistic human motion, since real human motion is captured, but is limited to motions that can be performed within the confines of a motion capture studio. Additionally, all of these methods are both complicated and time-consuming, making them fairly inflexible to the needs of animators.

The idea behind motion editing is to allow animators to alter the motions generated by these techniques to the requirements of each specific animation. For it to be used, the editing must be significantly simpler and easier than simply generating appropriate new motions via the original technique. Additionally, the edited motions must be as high-quality as the originals, and maintain any particular characteristics of the original that the animator might want; the simplest system in the world will not be used if it does not give good enough results.

As well, motion editing offers not only the flexibility to derive new motions from existing ones more easily than creating those motions from scratch, it also holds promise for generating motions that could not be created by the original method at all. Motion capture data, for example, is limited to what an available actor can do in the confines of a motion capture studio; any action that will not fit inside a studio or that a human can not do simply can not be captured, but perhaps can be created by editing motions that can be captured.

Similarly, motion editing may allow different generative techniques to be combined. The work of Unuma et al [13] suggests that the characteristic style of a motion can potentially be isolated and edited. This could allow the style of a motion captured actor to be altered to subtly change the motion or to make the motion fit more closely with other motions in the animation. As well, that style could perhaps be copied from motion capture data onto simulated motions, allowing, for example, keyframed motions to be used with captured motions in the same animation of a character without jarring differences between the parts of the animation which draw from different sources. Since humans are so skilled at interpreting human motion that it is

Operation	*Effect*
Blend	Generated motion combines the features of both sources into a single motion with the characteristics of each.
Overlay	Individual features of first motion are added to second motion without altering any disjoint features.
Transition	End of the first source motion is blended into beginning of the second motion to generate a single, longer motion.

Table 1: Fundamental motion editing operations

possible for a particular person to be recognized from motion capture data being performed by a simple skeleton after only a short time, or even from the motion of point lights on the ankles [3], it may be that these subtle characteristics will be necessary to take into consideration.

1.2 Motion Editing Operations

Motion editing can be thought of as a way to use generated motions as building blocks to create animations. The main operations available with these blocks are given in table 1.

1.2.1 Blending

Blending two or more motions allows interpolation between them, such as producing a medium kick from a low kick and a high kick, or extrapolation, such as producing an even higher kick. More generally, alteration of a single motion, such as forcing it to interpolate certain constraints, could also be classified as blending.

1.2.2 Overlaying

Overlaying takes two motions and causes the character to do some or all of each, resulting in a more complicated motion; for example, a kick and a martial arts block could be overlaid to create a motion where the character kicks while blocking without sacrificing the characteristics of either.

1.2.3 Transitioning

Finally, transitioning allows these motions to be strung together into arbitrary sequences of arbitrary length, changing a database of motion clips into an actual animation. It is worth noting that two main types of transitions exist; those between disjoint motions which are separated by some non-negative amount of time in the animation, and those between overlapping animations, where the character changes smoothly from executing one motion to executing the other.

1.2.4 Usage

These editing primitives offer tremendous flexibility in using motion clips to create animations or new motion clips. In theory, continued application of these primitives should allow a wide range of motions to be generated from a small starting set. In practice, motion editing methods often add noticeable artifacts to the resulting motion, limiting the scope in which they can be applied [1].

This raises the difficult question of how far motion editing can push the underlying motions and generation techniques while still producing good results. Unfortunately, this question has been essentially ignored to date and is far too large to answer in its entirety within the scope of this research; however, this research has taken an important first step by categorizing the quality of a test set of motions with an automatic metric suite. This categorization was compared with those generated by knowledgeable human viewers for validation. Further, since the test motions are automatically generated with no human intervention, information gleaned from this research will be directly applicable to the problem of attaining fully-automated generation of acceptable motions.

1.3 Metrics Overview

A metric is simply a method of measuring. Applied in this context, then, each metric one would apply to a motion is a quality metric, since the final value one wishes to optimize is the quality of the resulting motion. More specifically, however, any number of pieces of data could be measured in the

[1] An example of such a motion with an induced artifact is available at *www.cs.brown.edu/people/psar/motion.html*.

process of estimating the motion's quality. As well as directly examining the motion data - joint angles - metrics can measure derived quantities, such as joint torque, ground contact force, or balance. These measurements, in turn, are examined by the logic of the metric to form a quality measure, many of which are then combined to form the final quality estimate. Note also that more complicated metrics could examine the relationships between two or more data sources simultaneously and generate their quality measures accordingly.

Unfortunately, the previous work uses few or no metrics to measure the success of the editing method; energy minimization is almost uniformly the only metric employed beyond human intuition. While human intuition is a very powerful tool for evaluating motion that humans are familiar with, such reliance on it is limiting. Automatic animation generation from motion clips, for example, requires other metrics, simply because a human is not present. Additionally, it may at times be useful to have a more precise or quantitative measure of certain characteristics than human observation can provide.

Additionally, the wide range of potentially useful metrics can tax the capabilities of human observers to consider them all, even should they have the necessary expertise to do so. Physical metrics, such as musculoskeletal limitations or character balance, are perhaps the easiest to see since they directly affect the realism of the motion. Actor metrics are likely harder; these come from the actor's intent for the motion, such as speed, power, snap, accuracy, and so on for a martial arts kick. Finally, style metrics may be subtle enough that explicit human observation and evaluation is difficult. These metrics include motion appropriateness, which takes into account how a human catching a basketball will move differently than a human catching a bowling ball. Even if the basic motions are the same, the greater power requirements of the latter will move the action more into the large trunk muscles, producing subtle changes that will give cues about the weight of the object.

These considerations may be crucial for highly expressive or accurate animations. Teaching animations, for example, should usually be highly realistic so as not to mislead those trying to learn the skill; similarly, a system with sufficiently robust realism and actor metrics may be able to offer suggestions for optimizing those metrics that the human from which the motion was generated can then try to adopt to improve his or her performance.

Additionally, metrics would allow comparison of the results of different motion editing techniques, characterizing the strengths and weaknesses of

each approach and the degree to which a method can modify original motions while still producing output of sufficient quality for the application at hand.

1.4 Metric Classes

In the most general sense, there are three classes of motion metrics. In order from most general to least, these are:

1. Signal Processing Metrics
2. Assumption-Based Metrics
3. Task-Based Metrics

1.4.1 Signal Processing Metrics

The most general class of metrics are those based solely on processing of the motion signal with no a priori knowledge of the task being performed. For example, when dealing with motion capture data, no edited motion should, when its physics are examined, require joint torque spikes well outside the range that a real actor could possibly generate, regardless of the specific motion involved. Metrics of this type are valid on any motion, and are hence the easiest to use - they're always valid. As is often the case, however, these metrics are in many ways some of the least discriminatory available; the huge variability in real human motion precludes overly specific metrics from being generally applied.

1.4.2 Assumption-Based Metrics

Accordingly, the next class of metrics includes some assumptions about the motion under consideration, and each metric is invalid for motions that do not fulfill its assumptions. These assumptions can be very simple and general, such as that the character maintains contact with the ground at all times. Even an assumption as basic as this can both usefully audit the blending process and be invalid for whole classes of motions (such as jumps).

Other assumptions can be about the character of the motion. Many motions have a definite period of interest in which the most important activity, the task or action being performed, takes place. This being the crucial part of the motion, its existence and location provide useful information to a

blending system. For example, a motion clip of a martial arts kick has a definite action being performed - the kick - at a definite time. A feature detection metric that locates this kick can ensure any editing preserves this main feature; in particular, this is very valuable for ensuring that a transition automatically generated between two kicks actually has both kicks fully expressed in the resulting motion. Again, however, this assumption is invalid for some classes of motions - a "walking and looking" motion is unlikely to have any sharp increases in activity and so this metric is unlikely to provide useful information.

1.4.3 Task-Based Metrics

A more specific class of metrics is the set of those particular to a certain class of tasks or motions. Clearly, these are unlikely to be valid outside of their motion class; nevertheless, these metrics can be such powerful discriminators of motion quality as to be worth using. As an example, a large portion of martial arts kicks have the knee retract, extend to straight, and retract again, and detecting and using this pattern provides very valuable information. Since this metric is so simple, it can be very precise and usually provides a much better guide than more general metrics; moreover, the whole sequence of the kick is clearly delineated, allowing the system to easily tell which part of the motion must be preserved and where good places to transition into or away from the motion might be.

Fortunately, while these metrics should not be applied outside of their motion class, it is often obvious whether a source motion is indeed outside of that class. For example, while a front kick and a roundhouse kick would both have this characteristic knee motion, neither the "walking and looking" motion from the earlier example nor an axe kick (which is performed with the leg straight or nearly straight through the entire motion) would have the characteristic knee motion of a typical martial arts kick, meaning not only is it easy for the system to tell that this metric should not be used, but this also provides useful classifying information about the motion - it is not a typical martial arts kick. Indeed, this metric was tested in this research, so more details on this auto-applicability detection are provided in its description.

With sufficient classifying metrics, a sequence of conditional logic could be employed to classify or partially classify a source motion, automatically determining which assumption-based and task-based metrics are valid to use. With a system like this, only those metrics that could be confidently identified

as appropriate would be used, and a human user could optionally supply additional classification information to aid in metric selection.

1.5 Goal

The ultimate goal is to automatically generate edited motions from a broad range of sources, using a combination of these layers of metrics to ensure the results are of sufficiently high quality.

The goal for this research was to motivate more a more extensive examination of the possible roles and uses of metrics in evaluating and filtering edited motion with an informal pilot study. In particular, this research compares the subjective motion evaluations of human subjects with the computed motion evaluations of metrics to determine if there is indeed any useful correlation between the two.

2 Background

Motion editing as a whole has received a modest amount of attention. Michael Gleicher [4] [5] has used spacetime constraints, a local optimization method that applies results over the entire motion, to retarget motions to new characters and to apply constraints to those motions, although specifying high-level constraints in the complex mathematical methods he uses can be challenging [5]. Zoran Popovic [11] has worked with spacetime constraints and energy minimization to force motions to comply with constraints, such as alternative character or environmental parameters. Victor Zordan [15] has approached the problem using controllers which track and attempt to produce the original motion, subject both to certain requirements that the controllers must fulfill and to kinematic constraints used to preserve the sharpness of motion features such as impacts. Lee and Shin use hierarchical B-splines to preserve high-frequency information from the original motion while using spacetime constraints and joint-curve warping to alter the original motions [7]. Unuma et al [13] use fourier analysis to preserve the high-frequency data while isolating and modifying motion parameters such as gait.

The problem of transitioning between two motions has received less attention. Rose et al [12] use spacetime constrains and energy minimization to fill in the part of the animation between two disjoint motions in a manner somewhat similar to pose interpolation, while Lee and Shin [6] also make

some use of this technique. Unuma et al [13] transition between motions like walking and running by interpolating between the motion parameters, gradually changing the walk to a run. Witkin and Popovic [14] do simple joint angle blending, which Zordan [15] also makes some use of. This technique was also employed by Perlin [8] [9] in the Improv animation system.

However, the field of motion editing has been all but bereft of measurements of quality. All existing systems rely heavily on human intuition, either in their implementation or operation. Other than this, the only metrics used have been energy minimization such as used by Rose et al [12], which is likely not appropriate for energetic motions such as martial arts kicks, and simple process-checks such as examining a motion for ground penetration by the actor. While human intuition about motion is a very powerful tool, any fully-automated system is obviously unable to employ it, and the uses for such automatic systems are nontrivial - the video game industry, worth tens of billions of dollars annually, is only the most obvious application.

3 Approach

3.1 Core System

All motion files are stored in the Acclaim Motion Capture (.amc) format, which stores an Euler angle for each axis of each joint at each frame. Accordingly, taking all the frames one after another, each axis of each joint can be viewed as a signal of real numbers over time, and treated with signal-processing techniques accordingly.

In fact, the system used in this work deals primarily with these signals, and applies signal-processing techniques based on those of Bruderlin and Williams [1] to create the final motions. In particular, each signal is divided into frequency bands by progressively applying an expanding Gaussian kernel to the signal. Each kernel application acts as a low-pass filter, so a frequency band is simply the difference between the filtered signal and that same signal with the next level of low-pass filtering applied.

A single center of blending is chosen for the two input motions, as well as an offset for the start of the second input motion relative to the start of the first (see figure 1). An ease-in/ease-out blend is applied to each frequency band to transition from the first signal to the second, and, although the same center and offset are used for every signal and band, each level of

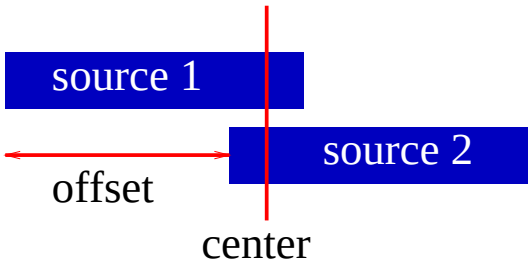

Figure 1: Blend parameters

frequency band in every signal has a particular blend radius around that center wherein the blend is effected, with the radius increasing for lower frequencies in accordance both with the observed frequency characteristics of the bands and with empirically determined principles.

Once the frequency bands have been transitioned together, the final transitioned signal is generated by adding together its constituent frequency bands, just as adding together the frequency bands of one of the source motions would exactly reconstruct that motion. As noted in both [1] and Burt and Adelson [2], this has the benefit of preserving information at all frequency levels.

3.2 Motion Creation

While the system was originally created to maximize the control over the resulting motion of a human animator working interactively, for this work the system was used to create transitioned motions automatically without human input.

1. Source selection

   Thirteen motion-captured single kicks were selected to be the source motions, and each edited motion consisted of two source motions transitioned together into a double kick. Each source motion was chosen uniformly, independently, and at random from the source set.

2. Motion positioning

   The system examined each of the two source motions to pinpoint the location of the kick impact in each, and then all pairs of integer center

and offset values were examined subject to the constraints that the parameters chosen must put the center of the blend at least five frames after the first kick and at least five frames before the second kick, as well as that the first and second motions must overlap by a non-negative amount. The center and offset pair which minimized the mass-weighted pose difference between the two motions in the center frame was selected, and then each was individually selected to be left alone half the time or permuted upwards by the square root of a number uniformly and randomly chosen from the set $[0, 200)$. This provides some variation in the resulting motions while biasing strongly towards the computed best parameters.

3. Motion alignment

   The root rotation of the first frame of the second source motion was then aligned to that of the first frame of the first motion, and the rotation required to effect this was applied to each frame of the second motion, effectively rotating the second motion to start out facing the same direction as the first. Similarly, the root translation of the center frame of the second motion was altered to put the left toe - most commonly the support foot for the kicks - in the same virtual world position as the left toe of the center frame of the first motion, with the required translation being applied to each frame of the second motion. While in theory either foot may be the support foot for kicking, the particular data set available for this study as well as the manners in which motions could reasonably combine were such that the support foot was always the left foot in the blend interval, and so no further distinction between the two is drawn in this analysis. In essence, the two motions were first pointed in the same direction and then aligned so the toe of the support foot was in the same place for each motion at the center of the blend. At this point, the blend was computed as described above and saved out to a file.

4. Corrective motion generation

   Since blending each band individually can lead to problematic foot motion, another blend with the same center and offset was made for the same source motions with the motions simply spliced together (effectively, a blend radius of zero, giving entirely the first motion before the center frame and entirely the second motion at and after that center

frame). Due to the toe alignment described above, this second blend has no discontinuity in the location of the left toe, and so provides a reference motion describing where the foot should ideally have been. This motion is used only to provide a desired foot location for the tracker to use when correcting the edited motion.

5. Dynamic tracking

   Both the edited motion and this reference motion are then fed into a dynamic tracker [10]. This tracker uses the edited motion to drive a character, subject both to realistic physics and the toe constraints set out in the reference motion, although the settings used for this work used quite liberal limits for joint torques and ground contact forces that could be applied. The output from this tracker, whatever motion the character managed to accomplish in this simulated physical environment, is the final result of the blending process.

3.3 Pilot Study

For the informal pilot study conducted for this work, thirty edited motions were made as detailed in the previous section. Each was rendered at 320 by 240 resolution using a fully shaded human character as shown in figure 2 [2]. Five human subjects, mostly with some familiarity with martial arts motions, were asked to skim over all of the motions and then go through them more slowly, marking each motion as one of Good, Mediocre, or Bad. The subjects were told that the task was to take two single kicks and create a double kick from those sources, and that both the quality of the blend and the overall realism and quality of the motion were important, and were then asked to categorize the thirty motions as one of Good, Mediocre, or Bad, with roughly equal numbers of motions in each category.

3.4 Metrics Used

A number of metrics were used to automatically compute, either singly or in combination, a quality value for the generated motions which would then be converted to the same Good, Mediocre, or Bad range as the humans used to facilitate comparison.

[2] A slideshow of all of the generated motions is available at *www.cs.brown.edu/people/psar/motion.html*.

Figure 2: A frame from one of the test motions

1. Contact Force Magnitude

   The character in the tracker had a weight of approximately 800 Newtons, so a contact force of 800 Newtons would be generated by the character standing quietly at normal gravity. The contact force divided by 800, then, is the multiple of its own weight that the character was exerting on the ground, meaning that very large numbers are unlikely to be suitable for a realistic motion, since no real human could provide those numbers.

Contact Force (N)	*Categorization*
< 1800	Good
$1800 - 3600$	Mediocre
> 3600	Bad

Table 2: Contact force metric thresholds

From real double kicks acquired by motion capture, it was determined that peak values of approximately 1200 N were typical for a real motion of this type. Due to a fairly high degree of stiffness in the tracker, it was known that such a stringent limit would rarely be met, so motions were marked as Good if they fell within 1.5 times this limit - 1800 N.

Any motions with values up to double this were marked as Mediocre, and the rest were deemed Bad motions.

2. Kick Counter

   In general, any region of very large and very rapid change in a joint angle parameter is apt to be an important feature of the motion and should most likely be preserved, regardless of the particular nature of the feature. This general principle is highly applicable in this particular domain, as kicks are usually easily detectable as such large and rapid changes.

Kicks in first source	*Kicks in second source*	*Kicks in edited motion*	*Same joint*	*Categorization*
N	M	$N + M$	Yes	Good
N	M	$\neq N + M$	Yes	Bad
-	-	-	No	-

Table 3: Kick counter metric algorithm

This metric counts the number of kicks in each of the source motions by finding the joint and axis with the largest and most rapid changes in Euler angle value and then counting the number of large acceleration spikes of approximately the same size in that signal. The metric counts the number of kicks in the edited motion in exactly the same way, and pronounces a motion as having passed if the edited motion has as many kicks as the two source motions summed together, or as having failed if it does not. Note that this metric is used as more of a filter than as a full categorizer, since it has such a limited scope of measurement that it is not suitable for categorization use alone. Further, not all kicks produce primary motion in the same joint, which the metric detects. Rather than attempting to compare motions across joints in such a manner, this metric detects if the source kicks have their primary motion in different joints and simply passes no judgement in that case.

3. Knee Retraction

   As was mentioned in the section on metrics, task knowledge can be used to exploit the very specific pattern of knee retraction and extension that characterizes many martial arts kicks. This metric searches each of the source motions for the characteristic extended-retracted-extended-retracted-extended pattern in the most active signal (in practice, the X axis of the kicking knee), and searches for the appropriate combined result in the edited motion, an extended-retracted-extended-retracted-extended-retracted-extended pattern. The edited motion is considered to have passed this filter if it has this double-kick pattern, and to have failed if it does not. Since not all kicks have this pattern, however, this metric is only applied if it detects this kick pattern in both source motions; otherwise, this filter is not appropriate and hence is not applied. Similar to the kick counter filter, this test is more appropriate for filtering bad motions out of a group than for categorizing all of them.

   It is worth noting that this metric is specific enough that it detects whether or not it is applicable. If it can not find the knee motions characteristic of this subset of kicks (including front, side, and roundhouse kicks, but excluding crescent, axe, and similar kicks, as well as virtually all non-kick motions) in the source motions, then it is clear that looking for those patterns in the edited motion is inappropriate - the edited motion will not add a kick that was not in the sources. Further, this lack of a kick could be taken as strong evidence that the source motion in question is not one of these types of kicks and this additional classificatory information could be used to further direct the categorization and filtration, but such conditional logic was beyond the scope of this study.

4. Support Ankle Torque Magnitude

   The support ankle is simply the ankle on the leg which is supporting the character during the kicking; as mentioned, in all cases in this study this was the left ankle. Since the way the blending process treats each frequency band separately can, with ill-matched source motions, create odd artifacts in the root rotation that the tracker can not entirely compensate for, this means that the final tracked motions sometimes have artifacts where the upper body moves around unnervingly while the support foot stays rooted solidly. That unnerving motion must be

Source 1	*Source 2*	*Blended*	*Categorization*
Yes	Yes	Twice	Good
Yes	Yes	$\neq$ twice	Bad
-	No	-	-
No	-	-	-

Table 4: Knee retraction metric algorithm

compensated for by the supporting ankle, so bad motions should show large spikes in the magnitude of the torque applied by it.

Examining the results of real motion-captured double kicks sent through the tracker showed that typical peaks in the support ankle torque magnitude were no more than 1100 Newton-Meters. Following the same reasoning as for contact force magnitude, approximately 50% leeway was allowed, meaning kicks with support ankle torques of 1600 Nm or less were marked as Good. In anticipation that these metrics would be used together and to obtain overlapping and non-identical coverage of the source motions, the threshold for considering motions Mediocre was set slightly higher, at triple the Good threshold or 4800 Nm, with all further motions being marked Bad.

Ankle torque (Nm)	*Categorization*
$<= 1600$	Good
$1600.1 - 4800$	Mediocre
> 4800	Bad

Table 5: Support ankle torque metric thresholds

5. Center of Mass Speed

   Unnervingly jerky motions also often create spikes in the speed of the center of mass of the character. As a final method of categorization, this metric was divided into categories based on the data; rough clustering

provided thresholds of under 0.60 m/s for Good motions and under 1.0 m/s for Mediocre motions.

Speed (m/s)	*Categorization*
< 0.6	Good
$0.6 - 0.99$	Mediocre
$>= 1.0$	Bad

Table 6: Center of mass speed metric thresholds

4 Results

Two of the main measures used in this section are *agreement* and *opposition*. One says two categorizers *agree* on the category of a particular motion if both of them assigned it the same category (ie. both assigned it Good, both assigned it Mediocre, or both assigned it Bad); hence, saying that 20/30 agree between the two categorizers means that for twenty of the motions, they assigned identical categories, whereas they differed for the other ten.

One says two categorizers are *opposed* on the category of a particular motion if one categorizer assigned the motion as Good and the other assigned it as Bad. If the categorizers did this twice across the thirty test motions, one would say they had 2/30 opposed.

Further, an important concept in this analysis is that of the human categorizations or "votes". Each human made 30 categorizations - cast 30 votes - and assigned one to each motion. When using a metric as a filter, then, one wishes to maximize the number of Good votes that are let through while minimizing the other votes that are not filtered out, particularly the Bad votes. This corresponds to maximizing the number of times a human will see a Good motion and minimizing the number of times a human will see a Bad motion, respectively, which is exactly the effect one wishes a motion-quality filter to have.

4.1 Human Results

Raw data for each subject is in table 15 in the appendix.

Each human-generated categorization was examined for agreement and opposition with the categorizations of all the other humans, with the average results shown in table 7.

Human	*Avg Agreement*	*Avg Opposed*
1	19.50/30	2.00/30
2	16.50/30	1.25/30
3	19.25/30	1.25/30
4	16.25/30	1.75/30
5	17.75/30	1.25/30
Average	17.90/30	1.50/30
Average (%)	59.7%	5.00%

Table 7: Test subject results and agreement

In general, the level of agreement between pairs of humans tended to be fairly consistent, with a low of 14/30 agreeing and a high of 23/30. The overall categorizations made by the test subjects are summarized in table 8.

Categorization	*Number*	*Percentage*
Good	52	34.7%
Mediocre	46	30.7%
Bad	52	34.7%
Total	150	100.0%

Table 8: Overall categorizations made by humans

4.2 Individual Metrics

Raw data for each metric is in table 16 in the appendix. A summary of the categorization and agreement results for each metric are in table 9, a summary of the results for each metric used to filter out Bad motions is in

	Contact Force	*Kick Counter*	*Knee Retraction*	*Support Ankle Torque*	*Center of Mass Speed*
Agreement with human	84/150 56.0%	36/90 40.0%	34/65 52.3%	76/150 50.7%	90/150 60.0%
Opposed to human	5/150 3.33%	26/90 29.0%	15/65 23.1%	7/150 4.67%	5/150 3.33%
Categorized Good	10/30 33.3%	13/18 72.2%	11/13 84.6%	10/30 33.3%	11/30 36.7%
Categorized Mediocre	8/30 26.7%	0/18 0.00%	0/13 0.00%	11/30 36.7%	8/30 26.7%
Categorized Bad	12/30 40.0%	5/18 27.8%	2/13 15.4%	9/30 30.0%	11/30 36.7%
Total Categorized	30/30 100.%	18/30 60.0%	13/30 43.3%	30/30 100.%	30/30 100.%

Table 9: Summary of metric categorization and agreement. Example: the contact force metric agreed with 84 of the total 150 categorizations made by humans, and gave opposite categorizations 5 times. It categorized 10 of the test motions as Good, 8 as Mediocre, and 12 as bad, thus categorizing all of the test motions.

table 10, and a summary of the results for each metric used to filter out non-Good motions is in table 11. This information is also presented in the text in the appropriate section.

When a metric is used as a filter, that means certain motions are omitted from consideration based on their categorization by the metric. For example, if a metric is used to filter out the Bad motions, only those motions it does not categorize as Bad will be considered for statistics leading from that.

- Contact Force

  With respect to human categorizations, 84/150 agree (56.0%) and 5/150 are opposed (3.33%). All motions were categorized, with 10 Good (33.3%), 8 Mediocre (26.7%), and 12 Bad (40.0%).

  Used to filter out Bad motions, this metric leaves 18 of the original 30 motions to consider. Of these 18 motions, the humans performed 90

	Contact Force	*Kick Counter*	*Knee Retraction*	*Support Ankle Torque*	*Center of Mass Speed*
Motions let through filter	18/30	25/30	28/30	21/30	19/30
Good votes cast to those motions	48/90 53.3%	46/125 36.8%	52/140 37.1%	49/105 46.7%	50/95 52.6%
Bad votes cast to those motions	17/90 19.0%	46/125 36.8%	46/140 32.9%	28/105 26.7%	18/95 18.9%
Total votes cast to those motions	90/150 60.0%	125/150 83.3%	140/150 93.3%	105/150 70.0%	95/150 63.3%

Table 10: Summary of results when metrics are used to filter out Bad motions. Example: 18 of the 30 test motions were judged as non-Bad by the contact force metric and hence passed through this filter. 90 categorizations were made by the human test subjects on these 18 motions selected by the contact force metric (one per person per motion), and 48 of those categorizations were Good (ie. votes cast for the motion seen being a Good one) while 17 were Bad. As this is a much better ratio than in the full set of motions, the contact force filter clearly biased the set it accepted towards Good motions.

	Contact Force	*Kick Counter*	*Knee Retraction*	*Support Ankle Torque*	*Center of Mass Speed*
Motions let through filter	10/30	13/30	11/30	10/30	11/30
Good votes cast to those motion	35/50 70.0%	30/65 46.2%	28/55 50.9%	35/50 70.0%	40/55 72.7%
Bad votes cast to those motions	3/50 6.00%	20/65 30.8%	15/55 27.3%	6/50 12.0%	3/55 5.45%
Total votes cast to those motions	50/150 33.3%	65/150 43.3%	55/150 36.7%	50/150 33.3%	55/150 36.7%

Table 11: Summary of results when metrics are used to filter out all non-Good motions. Example: 10 of the 30 test motions were judged as Good by the contact force metric and hence passed through this filter. 50 categorizations were made by the human test subjects on these 10 motions selected by the contact force metric (one per person per motion), and 35 of those categorizations were Good (ie. votes cast for the motion seen being a Good one) while 3 were Bad. As this is a much better ratio than in the full set of motions, the contact force filter clearly biased the set it accepted towards Good motions.

categorizations, of which 48/90 were Good (53.3%) and 17/90 (19.0%) were Bad, a marked improvement over the original distribution.

Used to filter out non-Good motions, this metric leaves 10 of the original 30 motions to consider, with 50 categorizations made by the humans. Of these, 35/50 (70.0%) were Good and only 3/50 (6.00%) were Bad, suggesting even this one metric could make an effective filter.

- Kick Counter

  This metric was applicable to 18 of the 30 motions, on which it gave a 36/90 (40.0%) agreement and 26/90 (29.0%) opposition to the human-derived categorizations. It categorized 13/18 motions as Good and 5/18 motions as Bad; as a pass/fail filter, it gave no Mediocre categorizations.

  Used to filter out Bad motions, this filter left 25 of the original 30 motions, on which 125 human categorizations were done. Of these, 46 were Good (36.8%) and 46 were Bad (36.8%).

  Used to filter out non-Good motions, this filter left 13 of the original 30 motions, on which 65 human categorizations were done. Of these, 30 were Good (46.2%) and 20 were Bad (30.8%).

  Results from this filter were disappointing, possible reasons for which will be touched on in the Discussion section.

- Knee Retraction

  This metric was applicable to 13 of the 30 motions, categorizing 11/13 as Good, 2/13 motions as Bad, and also by design gave no Mediocre categorizations. This had 34/65 (52.3%) agreement and 15/65 opposition (23.1%) to the human categorizations.

  Used to filter out Bad motions, this filter left 28 of the original 30 motions, on which 140 human categorizations were done. Of these, 52 were Good (37.1%) and 46 were Bad (32.9%).

  Used to filter out non-Good motions, this filter left 11 of the original 30 motions, on which 55 human categorizations were done. Of these, 28 were Good (50.9%) and 15 were Bad (27.3%).

  Results from this filter were disappointing, possible reasons for which will be touched on in the Discussion section.

- Support Ankle Torque Magnitude

  This metric categorized all motions, with 10 Good (33.3%), 11 Mediocre (36.7%), and 9 Bad (30.0%). This gave 76/150 agreement (50.7%) and 7/150 (4.67%) opposition to the human-derived categorizations.

  Used to filter out Bad motions, this metric left 21 of the original 30 motions, on which 105 human categorizations were done. Of these, 49 were Good (46.7%) and 28 were Bad (26.7%).

  Used to filter out non-Good motions, this metric left 10 of the original 30 motions, on which 50 human categorizations were done. Of these, 35 were Good (70.0%) and 6 were Bad (12.0%).

- Center of Mass Speed

  This metric categorized all motions, with 11 Good (36.7%), 8 Mediocre (26.7%), and 11 Bad (36.7%). This gave 90/150 agreement (60.0%) and 5/150 (3.33%) opposition to the human=derived categorizations.

  Used to filter out Bad motions, this metric left 19 of the original 30 motions, on which 95 human categorizations were done. Of these, 50 were Good (52.6%) and 18 were Bad (18.9%).

  Used to filter out non-Good motions, this metric left 11 of the original 30 motions, on which 55 human categorizations were done. Of these, 40 were Good (72.7%) and 3 were Bad (5.45%).

4.3 Metric Combinations

1. Kick counter and knee retraction

   Kick counting and knee retraction are relatively quick kinematic filters that can easily be applied while generating the motion in the blending system. Considering only those motions one or both marked as Good and none marked as Bad leaves 13 of the original 30 motions. Of the 65 human categorizations that represents, 30 were Good (46.2%) and 19 were Bad (29.2%), suggesting they are a somewhat useful online filter. This is combo 1 in table 12.

2. Contact force, kick counter, knee retraction

   In terms of stacking metrics, considering only those motions which were not categorized as Bad by any of the contact force magnitude, kick

	Combo 1	*Combo 2*	*Combo 3*	*Combo 4a*	*Combo 4b*
Motions	13	14	30	17	11
Categorizations	65	70	150	85	55
Good to	30	43	-	48	40
humans	46.2%	61.4%	-	56.5%	72.7%
Bad to	19	11	-	12	3
humans	29.2%	15.7%	-	14.1%	5.45%
Agreement	-	-	93	-	-
with humans	-	-	62.0%	-	
Opposition	-	-	5	-	-
vs. humans	-	-	3.33%	-	

Table 12: Summary of metric combination results

Good	*Mediocre*	*Bad*	*Category*
>= 3	<= 2	0	Good
0	<= 2	>= 3	Bad
other	other	other	Mediocre

Table 13: Consensus categorization algorithm

counting, or knee retraction metrics leaves 14 motions, 43/70 (61.4%) of the human categorizations of which were Good and only 11/70 (15.7%) were Bad. Comparing this to the results from using contact force magnitude to filter out Bad motions (53.3% Good and 19.0% Bad) shows the possible utility of stacking overlapping metrics. This is combo 2 in table 12.

3. Contact force, support ankle torque, COM speed - categorization

   In terms of categorization, using the three categorizers (contact force magnitude (CF), support ankle torque magnitude (SAT), center of mass speed (COMS)) with a consensus algorithm shows promise. The

three metrics have varying levels of agreement (17/30 between COMS and SAT, 24/30 between COMS and CF, and 21/30 between SAT and CF), but have no opposing categorizations. Since there are only three categories, there will thus always be a majority category given by the three metrics, and using this majority categorization gives 93/150 (62.0%) agreement and 5/150 (3.33%) opposition with the human-derived categorizations (which is, ironically, greater than the average agreement between the human categorizations themselves). This is combo 3 in table 12.

4. Contact force, support ankle torque, COM speed - filtration

   Moreover, using this consensus to filter out Bad motions leaves 17 of the original 30 motions, of which 48/85 (56.5%) were categorized as Good by the humans and 12/85 (14.1%) as Bad. This is combo 4a in table 12.

   Using it to filter out non-Good motions leaves 11 of the 30 test motions, with 40/55 (72.7%) categorizations of Good by the humans from this set and only 3/55 (5.45%) categorizations of Bad, showing that the consensus of these three metrics makes an effective filter to improve the subjective quality of motions to be shown to humans. This is combo 4b in table 12.

4.4 Discussion

4.4.1 Threshold Variations

While the threshold values for each of the metrics above were chosen carefully, these are by no means the only possible values. In particular, the contact force, support ankle torque, and center of mass speed metrics can be made into more or less discriminating filters by changing the threshold at which they reject a motion. This is a generalization of using the single categorizations given in the previous section to filter out either Bad or all non-Good motions, and simply has a metric filter out all motions whose values are above N, for some N that could either be set programmatically or by a user with a dial. This allows the metrics to easily vary their permissivity and adapt to different requirements. One application, for example, may have a need for motions of the highest quality - corresponding to restrictive settings of the metrics - while another application could be willing to sacrifice

some amount of motion for the increase in speed derived by being forced to reject less motions - corresponding to more permissive settings of the metric thresholds.

The results of altering the thresholds at which each of these three metrics accepts or rejects a motion are summarized in figures 3, 4, and 5. For each threshold setting, the number of Good categorizations by humans in the motions accepted by that threshold level is plotted, as well as the corresponding number of Bad categorizations, and the ratio of the two. As each metric is set to a more restrictive threshold level, both the Good and Bad values fall, but the Good more slowly, tending to cause a series of sharp increases in the ratio between the two. This suggests that using a restrictive threshold setting is highly beneficial. Potentially, using slightly less restrictive settings on several orthogonal filters would produce even better results; this seems to be a promising avenue of future work.

4.4.2 Underperforming Metrics

The task-based metrics of kick counting and knee retraction had surprisingly poor performance, especially considering their performance on earlier datasets. This may be in part an artifact of this particular dataset; the motion generation process randomly permuted the blend center and offset, but only ever increased either; this had the effect of making it extremely unlikely for either source kick to be accidentally removed from the resulting motion, which is what these metrics are designed to detect. While this would seem like a good approach to improving the quality of generated motions, such a bias creates other problems, such as late transitions, that were apparent to those viewing the motions.

4.4.3 Categorize for the Masses

The analyses presented use the data in its most raw form, in essence determining to what extent the metrics please all of the people all of the time. A more realistic goal is perhaps to attempt to please most of the people most of the time; this generates categorizations for the motions as shown in table 14.

With this approach, a motion must have been categorized as Good by a majority of the test subjects and not categorized as Bad by any of the test subjects for the final categorization to be Good, and vice versa.

Under this categorization scheme, 9 motions were categorized as Good

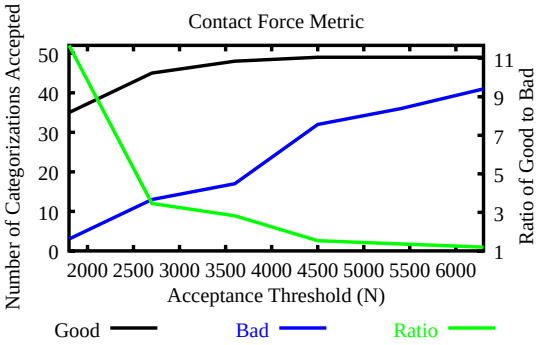

Figure 3: Contact Force

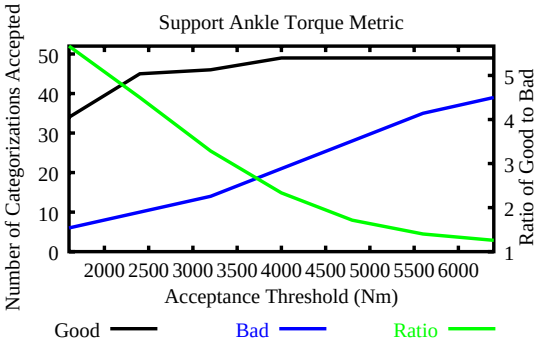

Figure 4: Support Ankle Torque

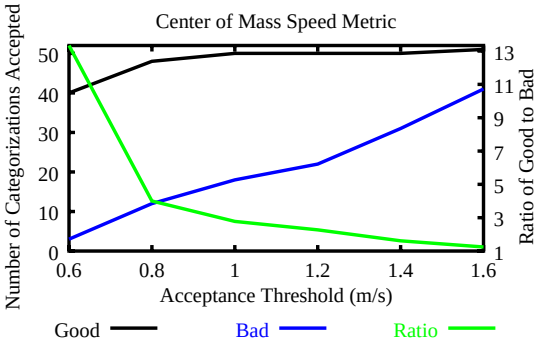

Figure 5: Center of Mass Speed

	Good	*Bad*
Total	9	7
Categorized correctly	8 88.9%	5 71.4%
Filtered Out	1 11.1%	7 100.%

Table 14: Consensus categorization results

and 7 were categorized as Bad by the human subjects. The consensus categorization algorithm (combination 3 in table 12) given in table 13 correctly identifies 8 of the 9 Good motions (88.9%) and 5 of the 7 Bad motions (71.4%). More importantly, however, none of those motions were given the opposite classification; as shown in table 14, using the algorithm to filter out all non-Good motions in a manner similar to combination 4b in table 12 and section 3 would remove all seven of these Bad motions, suggesting that while computable metrics may not be able to ensure no human viewer ever sees a motion he or she would consider Bad, they may be able to prevent any human viewer from seeing a motion that *most* people would consider Bad. Due to the vagarities of human intuition and preference, it may not be possible for metrics to please all of the people all of the time, but these results offer the hope that they may be able to please most of the people most of the time.

5 Conclusion

This work suggests that a small selection of well-motivated and powerful metrics can categorize generated motions as well as a human can in the sense that the automatic categorization will agree with a human's subjective categorization as well as another human's categorization would. Moreover, those same metrics can also be used as a powerful filter to automatically winnow out most motions which a human viewer would describe as of poor quality.

It is worth noting that, due to the highly subjective nature of human intu-

ition about motion, any future user study should include detailed guidelines on how to judge motions; indeed, discerning those guidelines may themselves be an area of fruitful study or collaboration with cognitive scientists. In particular, one issue raised by the subjects in this pilot study was that of judging a motion based on its desirability rather than its plausibility. The study guidelines must be clear whether the task is to judge a motion for how much it looks like something a skilled human *could* do or for how much it looks like something a skilled human *would* do - several of the randomly generated kick pairings struck some of the martial artists seeing them as more than a little strange.

Moreover, the poor performance of the task-based metrics compared to the success of the much more general assumption-based metrics was highly surprising. Although some reasons for this result have already been discussed, further examination of the roles of task-based and assumption-based metrics would be rewarding. In particular, these results are encouraging in that they suggest general metrics which could be applied to wide classes of motions can nevertheless powerfully discriminate between acceptable and unacceptable motions.

While this research conducted only an informal pilot study and a more rigorous examination of the agreement between computable metrics and human intuition is important for understanding the benefits metrics for motion editing can offer, the findings of this study are highly encouraging.

References

[1] A. Bruderlin and L. Williams. Motion signal processing. In *SIGGRAPH 95 Proceedings*, Annual Conference Series, pages 97–104. ACM SIGGRAPH, Addison Wesley, August 1995.

[2] P. J. Burt and E. Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 2(4):217–236, October 1983.

[3] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, 1977.

[4] M. Gleicher. Motion editing with spacetime constraints. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics*, pages 139–148, Providence, RI, April 1997.

[5] M. Gleicher. Retargetting motion to new characters. In *SIGGRAPH 98 Proceedings*, Annual Conference Series. ACM SIGGRAPH, ACM Press, August 1998.

[6] Sung Yong Shin Jehee Lee. Multiresolution motion analysis and synthesis. In *International Workshop on Human Modeling and Animation 2000 Proceedings*, June 2000.

[7] J. Lee and S. Y. Shin. A hierarchical approach to interactive motion editing for human-like figures. In *SIGGRAPH 99 Proceedings*, Annual Conference Series, pages 39–48. ACM SIGGRAPH, ACM Press, August 1999.

[8] K. Perlin. Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):5–15, March 1995.

[9] K. Perlin and A. Goldberg. Improv: A system for scripting interactive actors in virtual worlds. In *SIGGRAPH 96 Proceedings*, Annual Conference Series, pages 205–216. ACM SIGGRAPH, ACM Press, July 1996.

[10] N. S. Pollard and P. S. A. Reitsma. Animation of humanlike characters: Dynamic motion filtering with a physically plausible contact model, 2001.

[11] Z. Popović and A. Witkin. Physically-based motion transformation. In *SIGGRAPH 99 Proceedings*, Annual Conference Series. ACM SIGGRAPH, ACM Press, August 1999.

[12] C. F. Rose, M. F. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, September/October:32–40, 1998.

[13] M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *SIGGRAPH 95 Proceedings*, Annual

Conference Series, pages 91–96. ACM SIGGRAPH, Addison Wesley, August 1995.

[14] A. Witkin and Z. Popović. Motion warping. In *SIGGRAPH 95 Proceedings*, Annual Conference Series, pages 105–108. ACM SIGGRAPH, Addison Wesley, August 1995.

[15] V. B. Zordan and J. K. Hodgins. Tracking and modifying upper-body human motion data with dynamic simulation. In *Eurographics Workshop on Animation and Simulation '99*, Milan, Italy, September 1999.

Appendix A - Raw Classification Data

Motion	*Subject 1*	*Subject 2*	*Subject 3*	*Subject 4*	*Subject 5*
01	G	M	G	G	G
02	B	B	B	M	M
03	G	G	G	M	M
04	B	G	B	M	M
05	G	G	G	G	G
06	B	B	B	B	B
07	M	M	M	G	B
08	G	G	G	G	G
09	M	B	M	B	M
10	B	B	B	M	B
11	B	B	B	G	B
12	B	M	B	G	M
13	M	B	B	M	B
14	M	M	M	M	B
15	G	G	G	G	G
16	G	B	M	M	B
17	G	M	G	G	G
18	M	B	M	M	G
19	G	B	M	B	M
20	B	B	B	M	M
21	G	G	G	G	G
22	G	G	G	G	G
23	M	M	G	M	G
24	B	B	B	B	B
25	B	M	B	B	B
26	M	B	B	M	M
27	B	B	B	B	B
28	M	G	M	M	M
29	G	G	G	G	G
30	G	M	M	M	M

Table 15: Raw metric categorization results

Motion	*Contact Force*	*Kick Counter*	*Knee Retraction*	*Support Ankle Torque*	*Center of Mass Speed*
01	G	G	G	M	G
02	M	-	B	G	M
03	G	-	-	G	G
04	M	G	G	M	M
05	G	G	G	G	G
06	B	G	G	B	B
07	G	G	G	M	G
08	M	G	G	M	G
09	B	-	-	B	B
10	B	-	-	B	B
11	B	G	G	M	B
12	G	-	-	G	G
13	B	B	-	B	B
14	B	B	-	B	B
15	G	G	G	G	G
16	B	B	-	B	M
17	G	B	-	G	G
18	M	-	-	G	M
19	B	-	-	B	B
20	M	-	B	M	M
21	G	-	-	G	G
22	M	G	G	G	G
23	G	G	-	M	G
24	M	-	-	M	B
25	B	-	-	B	M
26	B	G	G	M	B
27	B	G	-	M	B
28	M	B	-	M	M
29	G	-	-	G	M
30	B	G	G	B	B

Table 16: Raw metric categorization results