Evaluating the perception of different matching strategies for time-coherent animations

Javier Villegas[a] and Ronak Etemadpour[b] and Angus Graeme Forbes[c]

[a]javier.villegasp@gmail.com, University of Arizona, Tucson, AZ, USA
[b]etemadp@okstate.edu, Oklahoma State University, Stillwater, OK, USA
[c]aforbes@uic.edu, University of Illinois at Chicago, Chicago, IL, USA

ABSTRACT

This paper introduces new terminology to describe the perceptual qualities of the non-photorealistic animation sequences created using an analysis/synthesis approach to rendering. Specifically, we propose the use of different matching optimization criteria as part of the creative control for generating animated sequences, or stylized video, and we explore the perceptual differences that are obtained when different optimization criteria are used. Additionally, metrics are introduced that describe the strengths and weakness of each of these matching strategies. Moreover, we show that these metrics may be useful for future evaluations of stylized video. We examine a series of sequences generated using different matching algorithms based on these metrics, and a user evaluation of 30 participants demonstrates that our objective metrics are perceptually relevant.

Keywords: Motion perception, interactivity, animation, temporal coherence

1. INTRODUCTION

Non-photorealistic animations, or stylized video, can be generated directly from video data. We explore an analysis/synthesis approach to the creation of these animations, where features are extracted from the video and information from those features is then used to create new graphic elements. However, analyzing video frames independently of each other can lead to non-continuity in the elements that form the resulting animation. In this paper, we propose and evaluate the use of combinatorial optimization algorithms to match the elements used for synthesis on each frame in order to reduce artifacts and to create animations that are perceived as temporally coherent.

Most image-based non-photorealistic techniques aim to imitate the appearance of hand-painted artistic effects, and an effective equivalent for moving images does not exist. Even if the problem of temporal coherence were to be solved, there are different alternatives for defining the way in which the elements should move and evolve. In this work, we introduce three definitions to describe the perceptual qualities of the sequences generated through our approach. The terms *smoothness*, *cohesion*, and *accuracy* are used to to assess the differences between various parameter matching algorithms.

Smoothness is a desirable quality of the animation that implies that the image elements should change slowly with time. Abrupt changes on position, size, shape, etc., are distracting and destroy the feeling of continuity. *Cohesion* is the spatial equivalent of smoothness. In a cohesive animation, features that are close to each other should behave in a similar way. *Accuracy* is an indication of how well the recreated image resembles the original analyzed picture. We introduce a set of metrics that quantify these terms and that can be applied to parametric spaces of any finite dimension. We then compare the objective description of the generated sequences (using these metrics) with an in-depth user study in order to demonstrate that these terms correlate to user perception.

A contribution of our work is the use of different matching optimization criteria as part of the creative control for generating stylized video, and we explore the perceptual differences that are obtained when different optimization criteria are used. Another contribution of our work is the introduction of metrics that can be used to describe the strengths and weakness of each of the matching strategies we explored. As the results of our user evaluation indicate, these metrics may be useful for future evaluation of stylized video.

2. RELATED WORK

Image-based non-photorealistic rendering (NPR) techniques have become increasingly sophisticated. However, there are difficult challenges that the discipline have yet to overcome.[1,2] One of them is the extension of the techniques to moving images. When frames are generated independently from one another the resulting NPR animation can have graphic elements that are not stable in time. They appear, disappear, or change too rapidly to convey the feeling of motion and continuity i.e., they do not have temporal-coherence. In their state-of-the-art review Bénard et al.[3] assert that it is the lack of temporal-coherence what has prevented NPR videos from adoption by the established media. Different authors have developed diverse strategies for the creation of temporal-coherent animations from video data. O'Donovan and Hertzmann developed a system that propagates paint strokes through time to create time-coherent painterly animations.[4] Collomosse et al.[5] describe a semi-automatic system that joins regions previously segmented on the frames of the input sequence to generate a region-based spatio-temporal representation of the video. This representation can then be used to generate various NPR effects. Our approach is also based on the matching of elements detected on independent frames, but, contrary to Collomosse et al., we propose the use of combinatorial optimization algorithms to perform the matching between elements, and we allow the use of any parametric element during the resynthesis.[6,7]

Another challenge for the NPR community (particularly for moving images) is the creation of methods for evaluating the results of the techniques. Some authors have focused on evaluating the performance of the participants in the realization of different tasks. Gooch et al.[8] designed experiments for evaluating the recognition and learning times of facial illustrations against photographs. Santella and DeCarlo[9] used an eye tracker to validate the effectiveness of abstraction algorithms in manipulating the areas of interest using different level of detail. Healey and Enns[10] asked participants to rank a series of images from multiple sources (including human and computer generated) according to different characteristics, such as artistic merit. Bénard et al.[11] describe the types of evaluations that have been used in the field and then correlates some objective metrics with the results of ranking experiments performed by humans on NPR textures. Part of the contribution of our work is the creation of metrics that can be used to describe the strengths and weakness of each of the matching strategies we explored and possibly in similar future works. We also performed user studies to validate our metrics with user's opinion. Section 4 outlines the metrics we created to evaluate the resultant animation and describes the user tests we performed. Section 5 shows the results of our metrics and explains how them correlate with the user studies.

3. MATCHING SYNTHESIS ELEMENTS

Our approach for generating time-coherent animations matches the parameters of particular features. A set of features that describe the image is calculated in the analysis stage. Each one of this features can be described with a finite set of parameters. The features on each frame are then matched with the features of the next frame. An assumption made throughout this process is that the number of detected features is the same throughout every frame in the video and that the matching can be posted as an assignment problem as follows:

- We want to minimize the sum of distances
$$\sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}X_{ij} \tag{1}$$
where C_{ij} are the elements of a cost matrix defined by
$$\mathbf{C}_{i,j} = d\left(\mathbf{z}_i[n], \mathbf{z}_j[n+1]\right). \tag{2}$$
Here $\mathbf{z}_i[n]$ is the feature i on frame n and $d()$ is the Euclidean distance. X_{ij} are the elements of an assignment matrix where $X_{ij} = 1$ if the feature i on frame n is matched to feature j on frame $n+1$ and $X_{ij} = 0$ otherwise.
- Finding the assignment matrix that minimizes the sum of distances between features is known as the linear assignment problem, and many algorithms for solving this problem efficiently can be found in the literature.[12]

- If the matching criterion is not to minimize the sum of distances but to match the elements in such a way that the maximum of all distances is minimized, the assignment problem is transformed to the bottleneck assignment problem. This problem is a variation of the original assignment problem for which, fortunately, efficient algorithms also exist.[12]

Once the matching is done for every feature on frame n, it is then known what parameters this feature will need to have in frame $n+1$. Each feature is then evolved towards its target state in the parameter space until one of two things happen: The feature accomplishes its final shape for frame $n+1$, or a new frame with a list of features is ready to be matched. The trajectory of the feature evolution within the parameters space and the number of intermediate frames that can be generated before a new target frame is input to the system are design parameters that significantly affect the final results. Those design decision imply a trade off between the *smoothness* of the animation and the *accuracy* of the representation. These concepts will be defined clearly in the next section.

Figure 1 shows the result of the assignment between two-dimensional features. Circles represent the features on frame n and stars are features on frame $n+1$. Every feature can be described by a vector of only two parameters (x and y position) and the matching between the objects is done using the minimum sum criterion on the left and the minimum maximum on the right. The cost matrix using Euclidean distance between elements is:

$$\mathbf{C}[n, n+1] = \begin{array}{cccc} 2.0000 & 2.8284 & 4.4721 & 6.0000 \\ 2.8284 & 2.0000 & 2.8284 & 4.0000 \\ 5.8310 & 1.4142 & 5.0990 & 4.2426 \\ 5.0000 & 5.0000 & 1.0000 & 2.2361 \end{array}$$

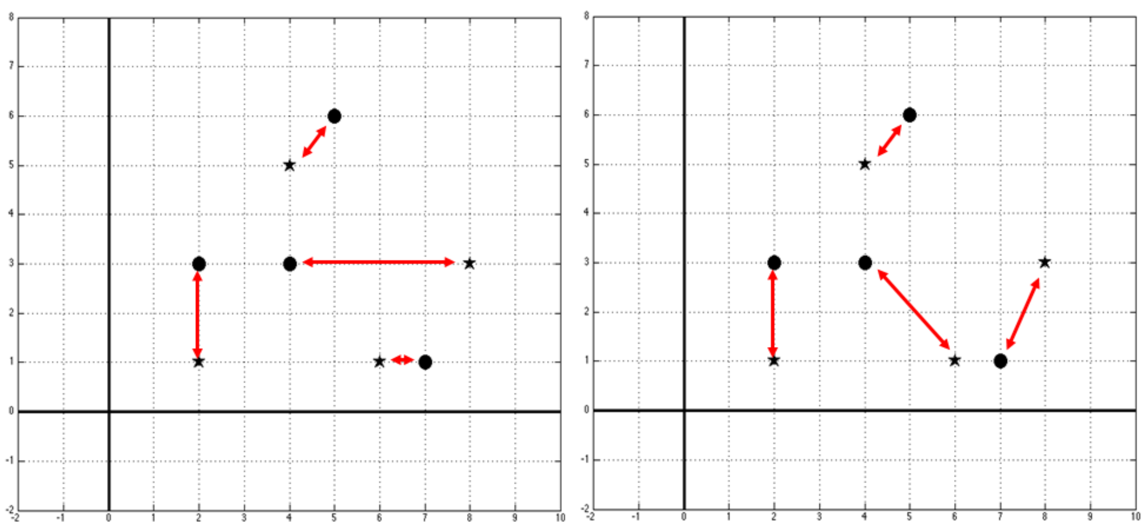

Figure 1. An example showing the two different criteria for doing the optimal assignment. Matching on the left minimizes the sum of distances while the matching on the right minimizes the maximum distance.

4. EVALUATING THE FEATURE MATCHING IN PARAMETER SPACE

By using different criteria with which to match the features in consecutive frames, our approach produces totally different visual output. We define three elements as descriptors of the quality of the resulting animation:

1. *Smoothness*, or "temporal continuity". Since the human visual system is very sensitive to abrupt changes, the elements of an image are expected to evolve continuously.[3] One definition of smoothness in trajectories

in different contexts is to use the minimum jerk rule.[13] Jerk is the third derivative of position, natural movements (e.g., hand movements) tend to be planned to minimize the jerk. A cost on the jerk of a trajectory can be defined as:

$$CJ = \sum_{n=1}^{N} \left(\dddot{x} \left[n\right]^2 + \dddot{y} \left[n\right]^2 \right) \tag{3}$$

Where: $\dot{x}\left[n\right] = x\left[n\right] - x\left[n-1\right]$; $\ddot{x}\left[n\right] = \dot{(\dot{x}\left[n\right])}$ and $\dddot{x}\left[n\right] = \dot{(\ddot{x}\left[n\right])}$. (A lower value on equation 3 means that the trajectory of a feature in its parameter space is smoother.)

2. *Cohesion*: This is the spatial equivalent of *smoothness*. If an animation is cohesive, then features that are close to each other in the parameter space should move in a similar fashion. Smith, et al., also states that a cohesive group motion is a desirable characteristic for computer generated mosaic animations.[14] To measure the amount of cohesion we define the "direction dispersion" (DD) as the averaged angle difference between a feature and the mean direction of its closest neighbors as:

$$DD = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \arccos \left(\left\langle V_{m,n}, \overline{V_{m,n}(d)} \right\rangle \right) \tag{4}$$

Where $V_{m,n}$ is the normalized direction vector of feature m in frame n and $\overline{V_{m,n}(d)}$ is the averaged normalized direction vector of the d closest feature to that feature. The *arccos*() function returns a always non-negative value between $[0, \pi]$ that represents the angle difference between the two vectors. This value is then averaged for all features M and all frames N. This metric can be used in parameter spaces of any finite dimension.

3. *Accuracy*: This property indicates how much the generated image in the animation diverges from the global shape it is trying to represent. The best way to calculate this representation error can be different for different types of features. In the next section we describe the creation of an artificial ground truth sequence that we use as reference for a particular selection of features.

4.1 Test Sequence

We created an artificial sequence of a ball made up of small circles. The ball is animated using time-dependent and space-dependent translations of the small circles. The resulting animation is a ball that is bouncing while rotating; its geometry also gets a little distorted, i.e., squished, at particular times during the bouncing cycle (see Figure 2).

Since the animation was created by small modifications of the parameters of the features (x and y position) we know a ground truth for the feature matching. We used the minimum sum and the minimum max criteria over each pair of frames of the sequence of the bouncing ball. We also tried a sub-optimal matching algorithm that for each feature on frame n assigns the closest available (i.e, not yet assigned) feature on frame $n+1$. Since the assignment results with this heuristic are strongly dependent on the starting point of the algorithm, we also tried a randomized version where the set of features is shuffled before the assignments are made. Figure 3 shows traces of the paths of a sample of elements for each of the four assignment algorithms (as well as the ground truth).

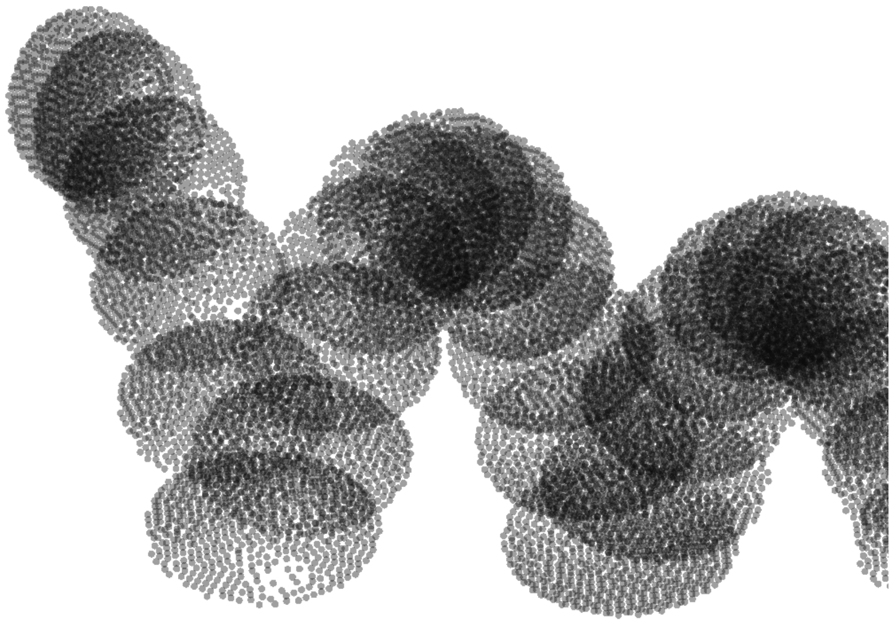

Figure 2. The test sequence created as a ground truth for the matching of the inner circles.

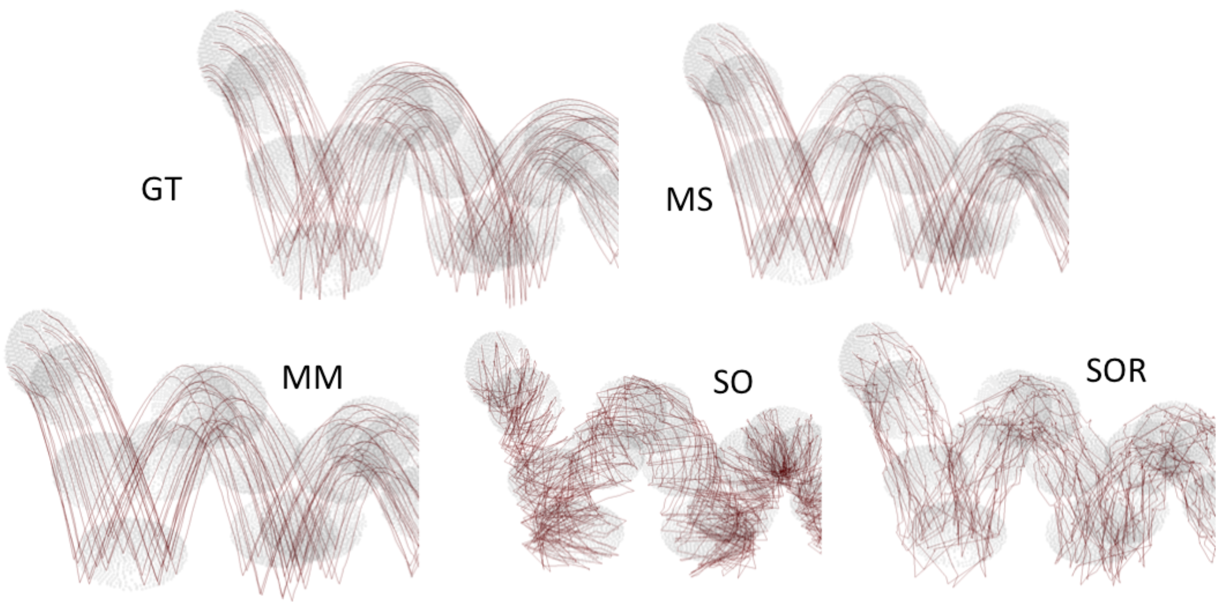

Figure 3. The bouncing ball sequence as generated with all the matching algorithms Ground Truth (GT), Minimum sum (MS), Minimum maximum (MM), First available (SO) and randomized first available (SOR).

5. MEASUREMENT RESULTS

Small changes to the parameters of each of the features in between frames are preferable over large changes. The different matching algorithms use different criteria to build the trajectories of the synthesis objects, thus the distribution of the amount of change is different in each case. If the features evolve at constant speed, larger jumps (changes in difference) also mean that some elements will need more time to evolve from one frame to the next, affecting the accuracy of the representation. Figure 4 shows the plot of the histogram of “jump sizes,” notice that the minimum sum (MS) criterion is accomplished by having a large amount of small jumps, at the cost of having some large jumps. With the minimum maximum (MM) optimization on the other hand, it is less likely to have large jumps, but it has a more sparse distribution of small jumps. Larger jumps are to be expected in either of the two versions of the sub-optimal matching algorithm, but in the randomized version the probability of having large jumps decreases faster.

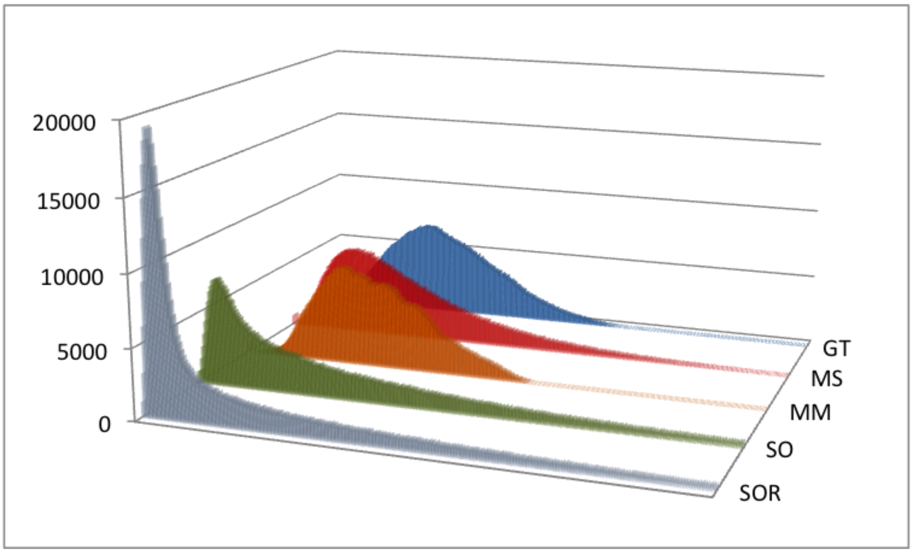

Figure 4. The histogram of jumps for each technique: Ground Truth (GT), Minimum sum (MS), Minimum maximum (MM), First available (SO) and randomized first available (SOR).

5.1 *Smoothness*

Using Equation 3 over all of the features provides us with a value indicating the smoothness of each alternative. Figure 5 shows a bar graph comparing the values obtained for each of the techniques. Ground truth, MS, and MM are calculated for two different numbers of intermediate frames in the bouncing ball sequence. The graph confirms that increasing the number of intermediate images improves the smoothness. It also shows that the MS and MM are comparatively similar to the ground truth with the MS performing always slightly better. The sub-optimal algorithms are both worse than the combinatorial optimization approaches, but randomization had an important impact in smoothing the result of the “closest available” heuristic.

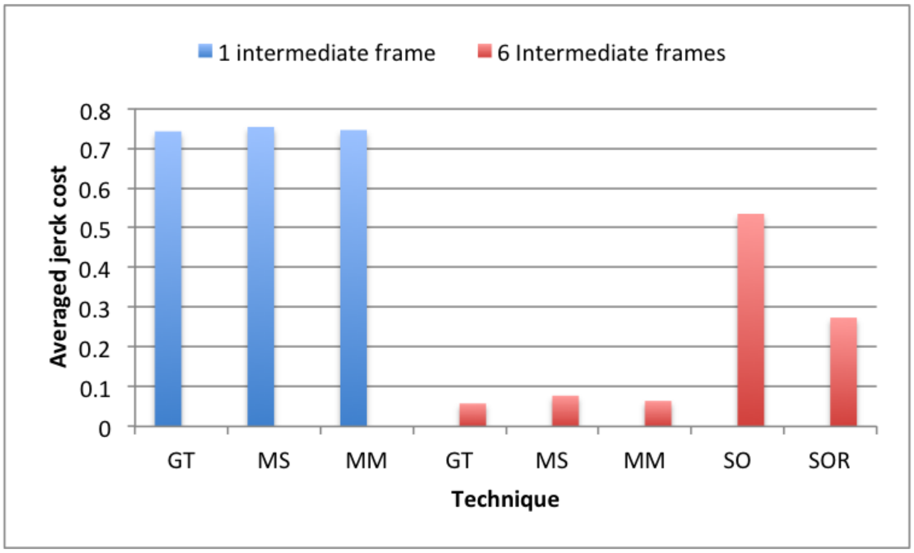

Figure 5. Comparing the smoothness value (averaged jerk cost) of each technique for two different number of intermediate images: Ground Truth (GT), Minimum sum (MS), Minimum maximum (MM), First available (SO) and randomized first available (SOR).

5.2 *Cohesion*

To compare the cohesion of the elements of each technique we calculated the direction dispersion (equation 4) for all the techniques. Figure 6 shows the results. The animation is more cohesive when the value is closer to zero, and less cohesive when it is closer to π. It is clear that the MS matching technique tends to generate more cohesive results than the others. It also makes sense that the value tends to increase when more intermediate frames are used since the features have more time to diverge to their destinies. The randomization stage of the sub-optimal algorithm radically worsens the value since it is less likely than a spatial pattern to have the assignment be reinforced on each frame.

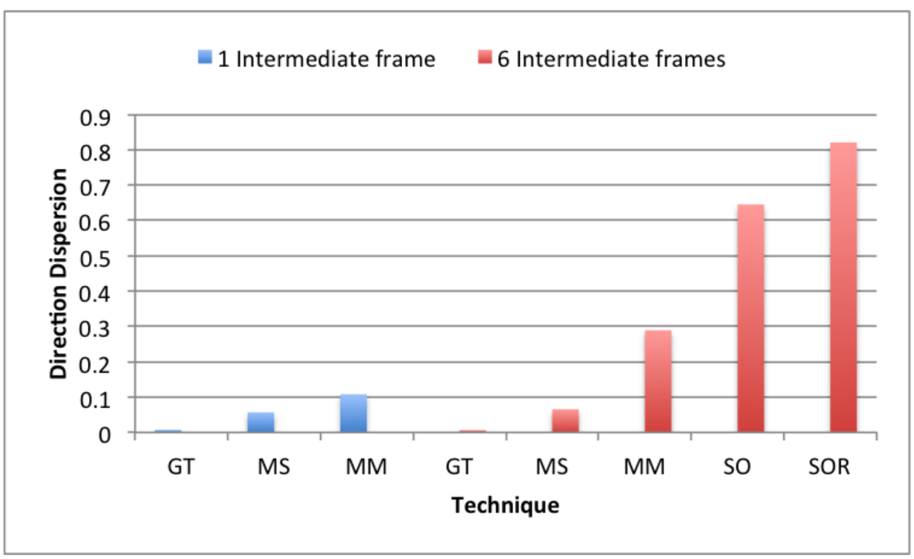

Figure 6. The direction dispersion for the same set of sequences than figure 5.

5.3 *Accuracy*

To measure how the accuracy of the representation is affected with each one of the techniques, we use the rendered binary images for the ground truth sequence and calculate the averaged pixel-by-pixel difference with the rendered version of each technique. The results are shown in figure 7.

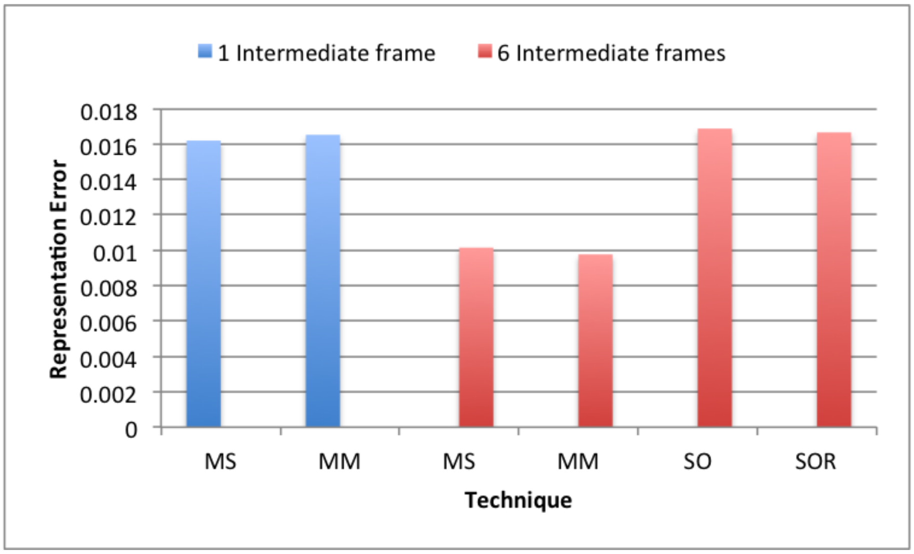

Figure 7. The representation error after rendering for the Minimum sum (MS), Minimum maximum (MM), First available (SO) and randomized first available (SOR).

It can be seen that, as expected, the representation error is lower when more intermediate frames are used, since the features have more time to evolve to their target positions. The graph also shows that the MS and MM techniques are comparable in terms of the accuracy of the representation, and they perform much better that the sub-optimal algorithms. There is no evident difference in the use of randomization in the sub-optimal first-available algorithms.

6. DESIGN OF USER STUDY

We created a user study in order to test whether or not the objective measurements indicated by these terms in fact correlate with user interpretation of these terms. Our user study had 30 participants, each of who was either an undergraduate or graduate student at the University of Arizona. We explained the concepts of smoothness, cohesion, and accuracy using a example animation. Then, in the first part of the test we showed them the bouncing ball animation and asked them to rate each of the three qualities on a five-step Likert scale. We showed them the animation at two different speeds by changing the number of intermediate frames but keeping the frame rate constant.

Several aspects were considered for the statistical analysis of this results. First, we compared the methods for each of the three defined concepts of smoothness, cohesion and accuracy by looking into the ranks reported by the subjects on each answer. Second, to verify if we found significant differences, we also performed pairwise comparisons of the groups to find the best and the worst results.

For all analyses, we computed means and standard deviation of the ranks. To test for statistical significance of the individual results, we first tested the distribution of the error values against normality using the Shapiro-Wilk tests. Since they had all non-normal distribution, we applied the Friedman test on K related samples when comparing more than two groups and the Wilcoxon test on non-parametric for two related samples when comparing two groups at the 0.05 level.

6.1 Smoothness Ratings

Although the average ratings of the two versions (fast and slow) of the ground truth (GT) sequences are always better than the others, this difference is only significant (i.e, worst case $p <= 0.00029$) for the slow sequence. The MM and MS algorithms are not rated in average significantly lower than the fast version of the GT sequence. On the contrary, the sub-optimal algorithms are always badly rated compared with the others (SO Slow vs MS Fast $p <= 0.000035$) and there is no a significant difference between the ratings of the randomized and the non-randomized versions. Figure 8 shows the results from the smoothness ratings.

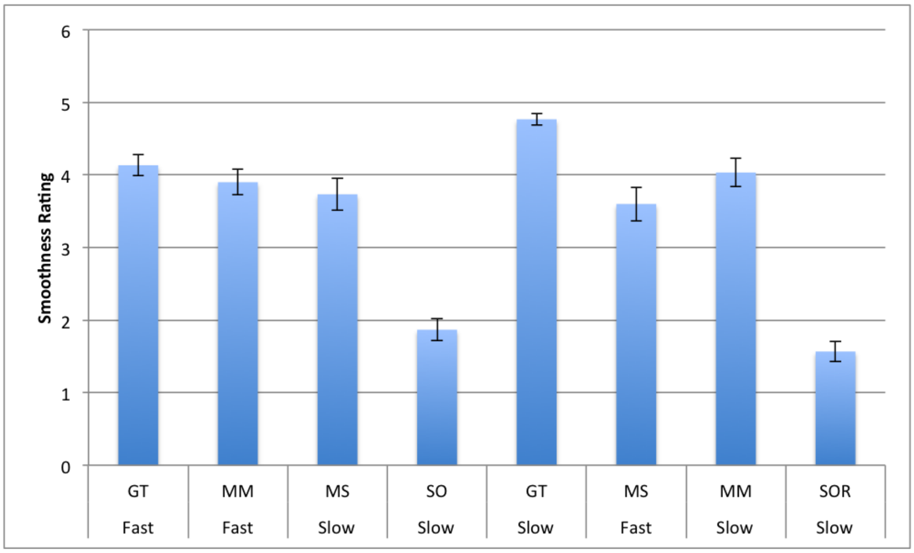

Figure 8. The averaged ratings for the smoothness of each technique in the same order they were presented to the participants: Ground Truth (GT), Minimum sum (MS), Minimum maximum (MM), First available (SO) and randomized first available (SOR). The slow version used six intermediate frames, the fast only one.

6.2 Cohesion Ratings

For the ratings of cohesion, differences of the reference sequence (GT) with the other are more important than with smoothness. Both versions of the GT sequence are always significantly better rated in average (i.e, GT Fast vs. MM Slow: $p <= 0.000085$, GT Slow vs. GT Fast: $p <= 0.046$). There are not significant differences in the ratings of the two versions of MM and MS. The sub-optimal algorithm without randomization was only rated lower than the GT version and the randomization was rated in average lower than all the other techniques and significantly lower (i.e. SOR Slow vs. MM Slow: $p <= 0.017$) than most of the others with the exception of the fast versions of MM and MS. The averaged ratings for the cohesion of each sequence are shown in Figure 9.

6.3 Accuracy Ratings

For this category the results were more diverse. The MM versions were particularly well rated, significantly better than the slow version of MS and the sub optimal algorithms. The randomized sub-optimal version was again the worse rated and significantly worst (i.e., SOR vs. MS Slow: $p <= 0.01$) compared to the rest of the techniques with the exception of the other sub-optimal algorithm. Figure 10 shows the averaged ratings for the accuracy of the different techniques.

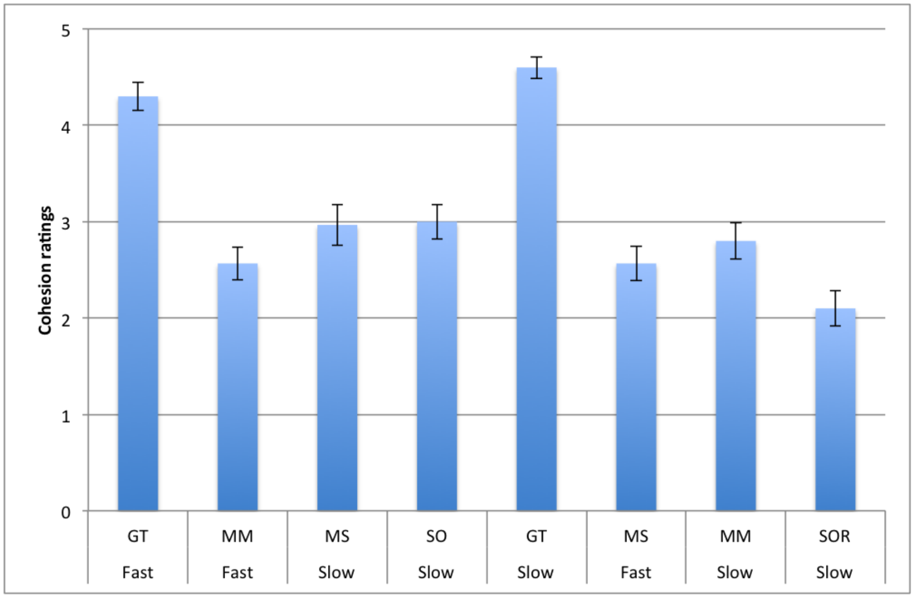

Figure 9. Averaged ratings for the cohesion of each technique.

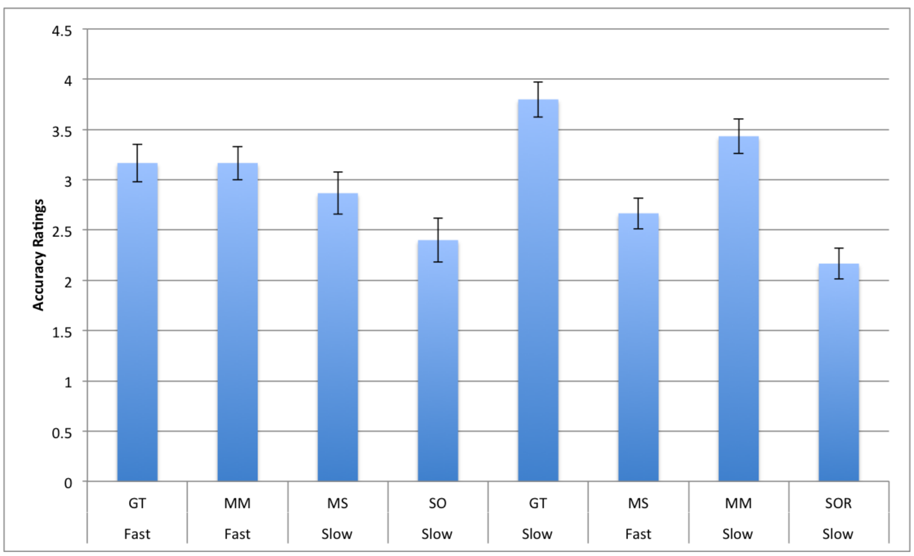

Figure 10. Ratings for the accuracy of all the techniques.

6.4 Pairwise Comparisons

In the second part of the test we presented to the participants sets of two animations side by side and asked them to chose the best one again in terms of smoothness, cohesion and accuracy. The results were conclusive; as expected the sub-optimal algorithms were always rated worse when paired against a combinatorial one in all three categories (SOR vs MM: largest $sig < 0.00035$ for cohesion). The MM was on average rated better than MS in the slow and fast versions (largest $sig < 0.016$ for cohesion). This result is clearly different from the single stimulus ratings where most of the time there was no significant difference between the combinatorial optimization algorithms. It is not surprising however that the qualities are perceived differently in a sequential or parallel comparison.

6.5 Objective and Subjective Measurements Comparison

Figure 11 shows the scatter plot of the subjective measurements against the results calculated from our metrics for the slow animations. It can be noted that in general a good rating average correspond to to a lower value in our metrics (in all our metrics lower values are better). This results gave a perceptual relevance to our metrics, since in general they produce a similar conclusion when analyzing these three qualities.

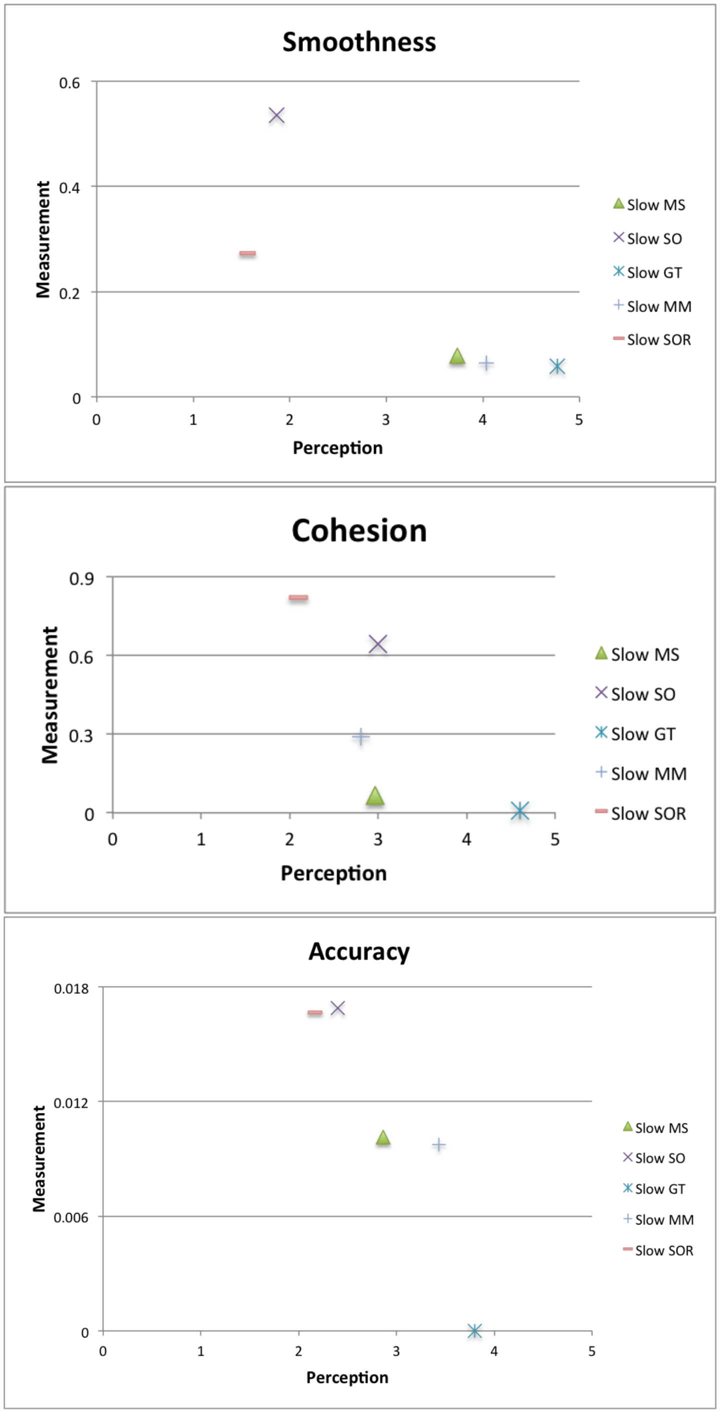

Figure 11. Comparing the objective and subjective measurements.

7. DISCUSSION OF RESULTS

Representing the synthesis elements (i.e., the feature vectors) as points in a parameter space helps us to extrapolate the problem of generating time-coherent animations to assignment problems and allows us to produce the metrics we described. This interpretation is not a universal solution for the creation of stylized videos but, as we showed in this paper, it is general enough to be used with different kinds of parametric features of any finite dimension.

One of the unique characteristics of our method is that, by knowing the evolution of the trajectory of a feature, we are able to apply different metrics to the resulting animation to evaluate its performance in terms of *smoothness* and *cohesion*. We also were able to generate a way to measure the accuracy for the test sequence we showed here, but that metric is not as universal as the other two and can vary depending on the features used. If we compare our metrics with the three goals defined by Bénard et al.[3]— *flatness*, *motion coherence* and *temporal continuity*— we can see that their term *temporal continuity* describes the same characteristic that we refer to here as *smoothness*. One of the novel contributions of our work is to propose the use of the *jerk cost* as a metric for *temporal continuity*. In section 6 we also showed that in our user studies, the *jerk cost* correlates with the subjective perception of smoothness (Figure 11). The term described as motion coherence by Bénard delineates something relatively similar to the spatial continuity that we measure and call *cohesion*. The difference is that for Bénard et al. a high value in *motion coherence* imply that the objects move no only together if they are close to each other but also in the direction of the container (the object they are shaping). A similar definition of that of Bénard et al. is also termed *cohesion* by Smith et al.[14] in their work on animated mosaics. However, our measure of *cohesion* is independent of the global motion of the container and gives an indication of in which degree the features behave similarly when they are close in the parameter space. Our direction dispersion equation is then another contribution to the field. Having objective metrics is important since in the future, algorithms that are optimal in terms of that metric can be designed. Another future work will be to generate a similar metrics that represent in what degree the objects are moving in the same direction of the container.

Our system was initially designed to work with the constraint that the number of features detected on each frame is constant. This apparently strong limitation can be overcome relatively simply in many cases. Many features vectors have one or more parameters that can be set in such a way that they are invisible at rendering. In this way "dummy" invisible elements can be defined in frames were few elements are detected. Another option is to repeat elements when the number of detected objects is low, so that in the rendering they will collapse to the same feature in the less populated frames. What elements should be repeated? Does it matter? For consistency with the optimization criterion the elements that have less average distance to all the target features should be the ones to be repeated. Similarly, for the minimum maximum criterion the features to be repeated would be the ones which have a lower maximum distance to all the target features.

We believe that in the same way that a filmmaker can choose a specific palette of colors to create a particular atmosphere, a computer animator should be able to choose a different matching strategy to generate different emotions in the audience. Our work is a first step in the understanding the perceptual differences of dissimilar matching strategies. Future work will further evaluate specific perceptual qualities of each of these matching algorithms.

REFERENCES

[1] Gooch, A. A., Long, J., Ji, L., Estey, A., and Gooch, B. S., "Viewing progress in non-photorealistic rendering through Heinlein's lens," in [*Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*], 165–171, ACM (2010).

[2] Kyprianidis, J., Collomosse, J., Wang, T., and Isenberg, T., "State of the art: A taxonomy of artistic stylization techniques for images and video," *Visualization and Computer Graphics, IEEE Transactions on* **19**(5), 866–885 (2013).

[3] Bénard, P., Bousseau, A., and Thollot, J., "State-of-the-art report on temporal coherence for stylized animations," in [*Computer Graphics Forum*], **30**(8), 2367–2386, Wiley Online Library (2011).

[4] O'Donovan, P. and Hertzmann, A., "Anipaint: Interactive painterly animation from video," *IEEE Transactions on Visualization and Computer Graphics* **18**(3), 475–487 (2012).

[5] Collomosse, J., Rowntree, D., and Hall, P., "Stroke surfaces: Temporally coherent artistic animations from video," *Visualization and Computer Graphics, IEEE Transactions on* **11**(5), 540–549 (2005).

[6] Villegas, J. and Forbes, A. G., "Analysis/synthesis approaches for creatively processing video signals," in [*Proceedings of the ACM International Conference on Multimedia (MM)*], 37–46 (November 2014).

[7] Villegas, J. and Forbes, A. G., "Interactive non-photorealistic video synthesis for artistic user experience on mobile devices," in [*Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*], (January 2014).

[8] Gooch, B., Reinhard, E., and Gooch, A., "Human facial illustrations: Creation and psychophysical evaluation," *ACM Transactions on Graphics* **23**, 27–44 (January 2004).

[9] Santella, A. and DeCarlo, D., "Visual interest and NPR: An evaluation and manifesto," in [*Proceedings of the 3rd International Symposium on Non-photorealistic Animation and Rendering*], *NPAR '04*, 71–150, ACM, New York, NY, USA (2004).

[10] Healey, C. and Enns, J., "Perception and painting: A search for effective, engaging visualizations," *Computer Graphics and Applications, IEEE* **22**(2), 10–15 (2002).

[11] Bénard, P., Thollot, J., and Sillion, F., "Quality assessment of fractalized npr textures: A perceptual objective metric," in [*Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*], *APGV '09*, 117–120, ACM, New York, NY, USA (2009).

[12] Burkard, R., Dell'Amico, M., and Martello, S., [*Assignment Problems*], Society for Industrial Mathematics (2009).

[13] Viviani, P. and Flash, T., "Minimum-jerk, two-thirds power law, and isochrony: Converging approaches to movement planning," *Journal of Experimental Psychology: Human Perception and Performance* **21**(1), 32 (1995).

[14] Smith, K., Liu, Y., and Klein, A., "Animosaics," in [*Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*], *SCA '05*, 201–208, ACM, New York, NY, USA (2005).