Animated People Textures

Bhrigu Celly and Victor B. Zordan
Riverside Graphic Lab
University of California, Riverside
{bcelly,vbz}@cs.ucr.edu
www.cs.ucr.edu/rgl

Abstract

This paper introduces a technique to create controllable animations of realistic figures of people starting from live-action video. The described synthesis of such 'people textures' extends previous work in video textures to allow the 'texturing' of human movement through human-specific feature extraction, coupled with careful *data mining*. In our approach, the video database is pre-processed to classify the motion of the human figures and identify the movements of repeated sequences using *data motifs*. Then, based on user input, novel sequences of video are computed with edits that are selected based on the raw footage found in the video database and performed based on morphing between segments to generate the transitions automatically. Applications for such animated people textures include video based animations for electronic games and creating background elements and special effects for movies.

Keywords: Computer Graphics, Three-Dimensional Graphics and Realism - Animation, Image Processing and Computer Vision, Applications

1 Introduction

We present an approach to control figures of people in video-based animation in order to create novel photorealistic sequences, based on pre-recorded footage. We anticipate that such *people textures* may be used to add directable animations of realistic human figures as background elements in movies and electronic games. Like other video texture approaches [1], we synthesize people textures from a database of video footage by rearranging the original frames of the source footage in the database. Our work extends video textures and previous, animated sprites [2] by investigating feature extraction and directed searches for good transitions specifically aimed for humanlike movement.

To generate animated people textures, we propose a dual search which first identifies gross likenesses between subsequences in the database and then performs focused queries to find good frame-to-frame transitions based on the initial search's result and user-inputted control. The final more precise transitions between sequences are created based on image-to-image comparison while the initial action recognition of the database is performed from features extracted from the data. The combination of these rough-to-fine searches can be used to create realistic novel motion video sequences of the figures recorded without exhaustive search over all frame-to-frame comparisons.

To minimize the need for computing frame-to-frame comparisons, trajectories of motion features derived from the images are compared using a stochastic *motifs* search. The term motif is coined from bio-informatics and refers to a repeated subsequence, like those found in DNA for example. The motion of the actor is divided into a set of such repeated subsequences based on matching patterns found in the features extracted from the source footage. From

these different subsequences found by the motif search, a much narrower search for frame-to-frame comparisons are made to find good transitions. Thus, the motifs provide a good guess for where the smooth transitions may be found, while also providing an *action alphabet* for high-level control. The animated people textures can then be found randomly, or based on control direction from an animator who selects the desired sequences of actions from the discovered motifs. Transitions finally smoothly morph between the raw segments. This paper demonstrates the power of this approach with two examples, the first from male and female actors performing yoga *asana* sequences and the second of an actor performing martial arts *kata* moves.

1.1 Related work

Video textures is a relatively new area of research in computer graphics, stemming from the seminal work by Schödl et al in 2000 [1]. Since, Schödl and colleagues have offered several related follow-up papers [3, 2, 4] and the work has spurred a new area of related research in the reuse of motion capture for animation. Video textures are continuous streams of video footage that are created from the reordered frames of raw footage collected apriori. By selecting good transitions between frames (edits) the resulting novel footage appears to belong to a continuous, non-repeating stream of the original content. Their work demonstrated the technique on several examples including a candle flame flickering, balloons floating in the wind, and children swinging back and forth on a swingset. Follow-up work in video textures offered improvements for finding and evaluating good transitions [4] and means for controlling sequences using machine learning [3, 2].

In a related effort, re-use of motion capture was similarly treated where several researcher offer solutions for fining good transitions between motion capture samples to create longer continuous streams of motion data [5, 6, 7, 8, 9]. Unlike the seaming of two images, the transitions of motion capture clips provides a higher degree of control over the final motion because there are many fewer parameters to match, and the data is much more structured. To be clear, consider a good transition between two n x m-sized images requires n x m pixels be matched in the final sequence (and even more when considering the additional frames in the edit,) yielding 307200 matched elements in standard video, whereas the usual 50-75 degree-of-freedom motion capture-driven character yields as many matched elements in the final motion sequence. Further, error associated with less-than-perfect transitions can by blended more easily in motion capture because the matched elements have structural meaning, ie. a specific degree of freedom in the character's skeleton.

People textures offer a solution between the general use of video textures and the re-editing of human motion capture data by exploring the re-ordering of human motion found in video and exploiting the specific qualities inherent in video footage of human motion. We manage the large number of pixel-based image matches in two ways: first, by performing a fast search on high-level features, we narrow and limit the image-to-image searches and, second, by green-screening and careful pre-segmentation we minimize the pixels to be matched. Like Schödl and Essa's work in controlling video sprites, we focus our efforts on maintaining high-level control over the resulting footage, but using a novel two-tier solution which separates the control from the selection of transitions. Further, while Schödl and Essa set forth to control their video sprites to follow paths (while also avoiding collisions and adding timing constraints,) we aim for a 'director-level' control of the people textures which may be controlled instead with action commands like "perform a high kick" or "transition to a downward-facing dog" pose in yoga. In this way, our efforts begin to approach the specialized directability of some speech-driven video face systems [10, 11] although our approach and the domain are quite different.

2 Human segmentation and analysis

For the testbed of this paper, we focus on animating the figure of a single person performing well-defined actions. The individual is recorded against a green screen so that s/he can be easily separated from the background (and subse-

quently, be trivially added to other video sequences using compositing software). The actor is recorded performing several combinations of the predefined sets of actions. The video is then analyzed as a continuous stream of images and the repeated segments are classified as actions or a set of actions automatically.

2.1 People extraction

The actor's body is first separated from the background in the video through image processing using either background subtraction or color tracking [12]. The result is converted to black and white by thresholding and the final binary image is used to compute features about the motion of the actor. Such features are then used to classify the actions in the recorded video. To assess the utility of different features for matching people textures, we investigated several metrics including the area, height, and width of the segmented foreground image (the human) as well as its bounding box area and aspect ratio. Through experimentation, we found that the time trajectory for the aspect ratio of the sub-image's bounding box was a useful single-data series, likely in part because it embedded information about motion in both the vertical and horizontal directions and inherently removed scaling artifacts created as the actors appeared forward and aft along the direction of the camera's viewing axis. These steps are high-lighted in Figure 1.

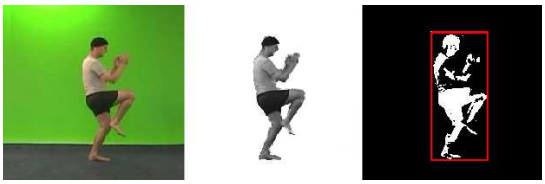

Figure 1: Stages of image processing from raw frame, to grayscale with background subtracted, to threshold and with bounding box. Bounding box aspect ratio and centroid were both used in identifying repeated actions in the video.

2.2 Finding motifs

Searching for data motifs was originally adapted from computational biology as a means of finding repeating patterns within sequential time series [13, 14]. The advantage of motifs over other pattern matching techniques is that the approach does not require apriori templates or example sequences. Instead, the motif search finds the K most-often repeating, non-trivial, mutually-exclusive repeating patterns present in the data series. To perform a direct motif search on a given time series, the user must define the length of the desired subsequence, n, and the error tolerance, range R, which is equivalent to the sum of the Euclidian distance (ED) between a tested subsequence and proposed exemplar. Then, the most significant motif is defined to be the subsequence C_1 that has the highest count of non-trivial matches (error less than R) and the i-th motif up to K is the next most significant sequence C_i after the $(i-1)$-th motif, given that the distance between it and all previous motifs affords its uniqueness (or $ED(C_i, C_j) > 2R$ where $1 \leq j \leq i$.) Direct exhaustive search for the set of motifs is easily implemented by comparing all subsequences of length n.

Because the exhaustive search for such motifs is computationally expensive, Chiu and colleagues propose a stochastic approximation which supports faster searches and they describe an iterative algorithm to find motifs [14]. In their algorithm, the time-series of interest (the aspect ratio of the bounding box over time in our case) is transformed into a compact symbolic representation which affords dimensionality reduction and *lower bounded* search. Then, fast comparisons are performed in the reduced space using a technique called random projection. We refer the reader to this work for further details on their stochastic implementation. For the purposes of finding people textures, we can regard their motif search as a black-box system which supplies matches from the given time series with successive refinement and which may be halted when a satisfactory solution is found.

For the motif discovery needed for video textures, we select the top K motifs discovered after a few iteration of the described algorithm and assign each as letters in the *action alphabet.* We experimented with several different feature extraction metrics and a number of lengths, corresponding to 40, 80, and 160 frame-length sequences which seemed appropriate for the frequency of actions in our video databases. Two iterations of the motif-discovered transition ma-

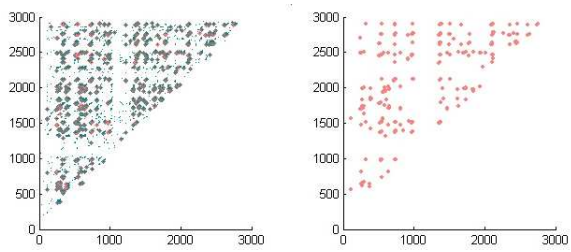

Figure 2: The motif-transition matrix at intermediate and final iterations. The ordered vertical and horizontal repetition is indicative of the multiple matching sequences associated with single motifs.

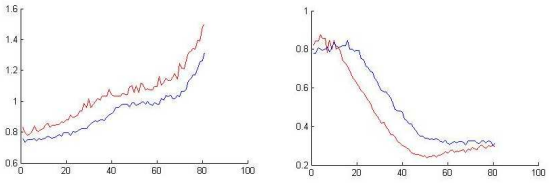

Figure 3: The bounding-box plots of two motif examples found for the female yoga data over 80 frames. Each plot shows the similarity of two of the matches found for the different motifs.

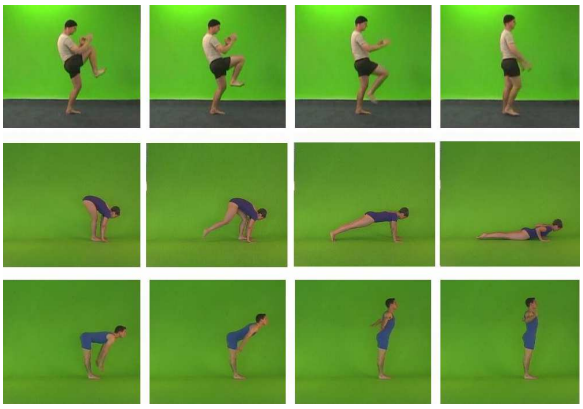

Figure 4: Corresponding footage for one motif action from the *alphabets* of each database. Such actions allow the user to assemble different sequences that transition through these actions.

trices for the female yoga database appear in Figure 2. In this figure, the colored areas indicate good starting points for potentially matches. As the iterations progress, more refined matching is determined and after halting, corresponding results from the motif analysis are shown in Figure 3 and Figure 4.

3 Rules for animating

During the video playback when a motif is active, the transition matrix provides a number of possible forks in the video sequence. If the animator selects a new fork, the algorithm compares raw sequences to find the best place for the switch and generates a smooth transition surrounding this point by morphing between the images. A schematic for this step is shown in Figure 5.

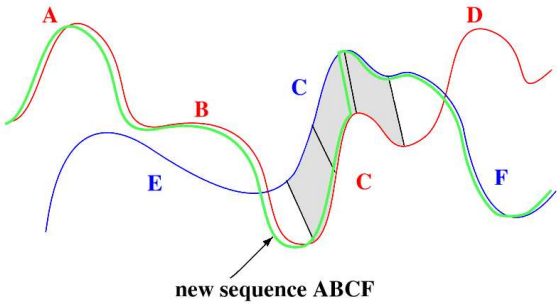

Figure 5: Schematic for using motif *alphabet* to generate novel sequences. The motif C shown is aligned between two raw sequences. The thicker, green path indicates the path for the novel sequence which follows a smooth transition between the raw examples selected based on frame-to-frame errors.

3.1 Silhouette-to-silhouette comparison and selection

To create a transition for motif C_κ, we look at the series of frames associated with $C_{\kappa\alpha}$ from the current video playback to $C_{\kappa\beta}$ the next (selected) sequence. For comparison, the algorithm uses the sub-images of the original found by removing the background (see the middle image of Figure 1.) We call this image a "silhouette." The sequences of interest are then

defined by the series of silhouettes in each:

$$C_{\kappa\alpha} = a_1, a_2, \ldots\ldots a_n$$
$$C_{\kappa\beta} = b_1, b_2, \ldots\ldots b_n$$

The system builds a silhouette-to-silhouette comparison matrix, assessing the error between samples across the selected motif sequences. The comparison matrix is computed to include a measure of similarity between all combinations of silhouettes from the raw sequences. For the image similarity measure of two frames a_f and b_g, the system first aligns the centroid and scales the size of the bounding box of the second, creating b'_g. Then it computes the error $\xi_{f,g}$ as in the original video textures work to compute the image difference. Thus, for a_f and b'_g the system computes

$$\xi_{f,g} = (a_f - b'_g)^2$$

as the silhouette-to-silhouette distance and stores this value in the comparison matrix. We define the smallest value in the matrix as the two silhouettes that are the closest to each other in the two raw sequences. Like the original video textures approach, the two closest consecutive frame matches are selected for the transition point in order to preserve the dynamic information in the generated texture. We add the values in the comparison matrix to compute the error for a pair of consecutive frames and select the pair with the minimum difference.

3.2 Transition using feature-based morphing

To seam the join smoothly, we apply a feature based morph over the selected frames following the approach described by Beier and Neely [15]. Feature based morphing smoothly transforms the source image to the target image based on a defined set of feature lines. Diverging from Schödl's approach, we use morphing for animating people textures as we found that simple cross-dissolves did not tend to work well and the feature based morphing provided more flexibility and control compared to other techniques. For our purposes, we use a short, fixed length transition where the source images are distorted to reach the target images across the transition. The weights and an example morph resulting from our method appears in Figure 6.

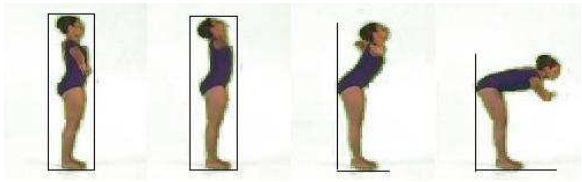

Figure 7: Feature lines for the actress when is almost stationary for the two frames on the left and when she is in motion on the right. Note, when in motion, the side near the back and the feet are recognized as stationary.

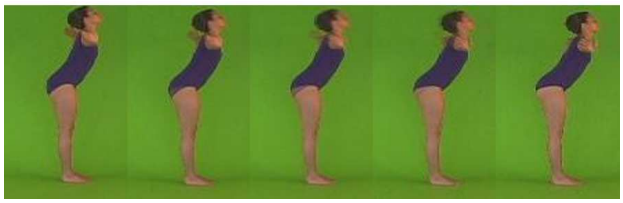

Figure 8: Close-up of morphed transition. In the middle frames, note problem areas like the bodysuit and the far hand as they disappear across the transition.

For the morph features, we use a set of lines on the source image and the destination image corresponding to fixed horizontal and vertical edges of the silhouette bounding boxes for the selected frames. To assess which edges of the bounding box are fixed, the algorithm evaluates the location change of the edges over an short interval (± 5 frames) surrounding the two frames. Edges which move very little (under a set threshold) are assumed fixed and used as features in the morph. Work by Liu et al. in patterned-based texture morphing inspired this feature-selection approach [16]. The edge features extracted for two cases are shown in Figure 7. In general, we found this method produced smooth movement and avoided sudden changes (shocks) which can easily be detected by the eye. A close up of an example morph shows more detail in Figure 8.

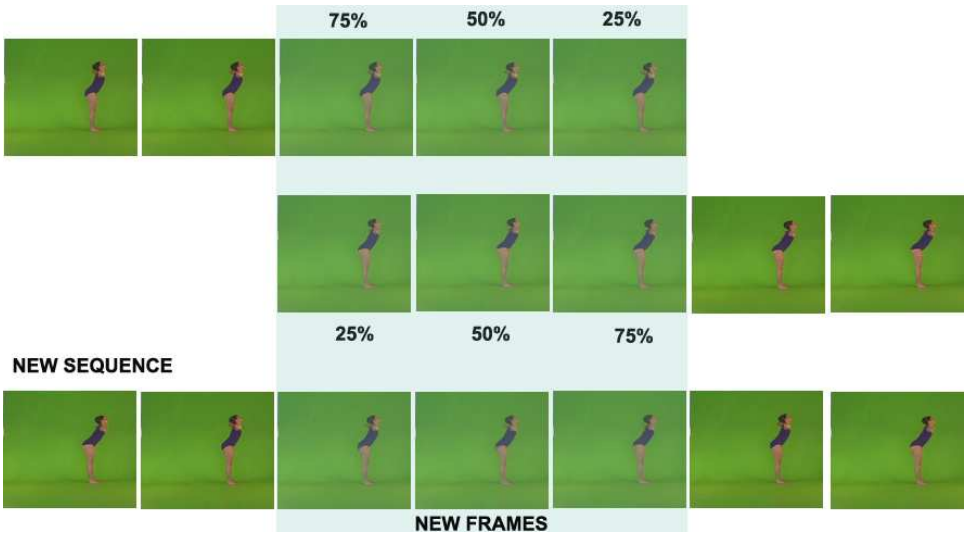

Figure 6: The bottom row shows a sequence of transition frames with the middle frames (shaded) generated using morphing. The values show the weighting percentages.

4 Implementation details and results

For our experiments we used a Canon ZR40 camcorder with the shutter at 1/60, video size of 720x480 pixels, and captured video at 30 frames per second. The accompanying video clips demonstrate several textures produced by the methods described, for three test cases. **Yoga Male** - 7500 frames in which the actor performed two yoga asana (pose) sequences several times each. We created a 75-second movie clip of the actor performing reordered asanas. There are four transitions in the sequence, including examples both while he is moving and while he is stationary. **Yoga Female** - 9000 frames, same as male version. We created a 35-second video clip of the female with four transitions in the video including motion in which she goes from standing to prone and back. **Martial Arts** - 9000 frames of actor practicing katas - a sequence of stances and kicks. We generate a novel 75-second video clip with four transitions in the video. All transitions were generated automatically by our algorithm.

5 Conclusions and Discussion

An important feature of people textures is the initial motif search which identifies key repeated actions in the source footage. Unlike the less structured footage of hamsters, flies, and fish analyzed by Schödl and Essa in their work on controlling video sprites, human motion contains immense structure, based in part on the intention in the tasks performed. In the proposed approach, we exploit this characteristic specific to people and their movements. While we expressly selected the testbeds in this motion because of their formal and repeatable nature, more casual motions, such as those found in the buzz of movement in public squares, also contain much more structure than seen in the animal motions described previously. And so, we predict that an algorithm which exploits the structure of human motion will inevitably be used to create video-based (and possibly, motion capture-based) animations containing people.

We anticipate that several improvements could be made in the creation of video people textures. In morphing, we found that a small number of edges as described was sufficient but it is likely that more features could result in better morphs. For example, the Beier and Neely work use around 10 features to morph the human face in their examples. (However, in their work, features were selected by hand and in ours feature selection is automatic). Likewise, depending on the footage, different colored regions could be exploited during feature extraction for morphs. We experimented with morphing with features related to the woman's purple

bodysuit, for example, and our preliminary results seemed promising. The head of the actor performing the yoga sequences caused problems when creating the transition. While the action of the actor after the transition remained contiguous, the head motion included a noticeable disturbance. Detecting and solving this problem still remains. Also, when the actor performing the martial arts sequence was using his left and right leg for the same kick, the legs look visually similar to the algorithm and it sometimes found false matches. We feel higher resolution video with good lighting may solve this problem.

The animated people textures resulting from our two-tier motif-based approach offer a strong indication that the paradigm of video texturing may indeed be applied to create controllable photorealistic human sprites. In the future, we aim to remove limitations associated with the people textures presented in several ways. First, we want to improve our existing algorithm to use more feature lines on the actors body during transitions and would like to use more information to drive motifs then the bounding box aspect ratio. In general, we would like to analyze footage from more natural settings without the use of controlled lighting and environments. Also, while we limit our investigation to a single person, the search for people textures with many individuals is certainly appealing, especially for the applications associated with controllable background extras for movies. Finally, we envision that the use of multiple-cameras approaches like [17] and [18] may be merged with video texturing-type methods and result in life-like, fully 3D controllable people textures following the recent re-use trends seen for editing human motion capture.

Acknowledgements

We take this opportunity to thank Matt Fast, Selena Brown and Erik Guzman who were the actors for our experiments.

References

[1] Arno Schödl, Richard Szeliski, David H. Salesin, and Irfan Essa. Video textures. In *Proceedings of ACM SIGGRAPH 2000*, Computer Graphics Proceedings, Annual Conference Series, pages 489–498, July 2000.

[2] Arno Schödl and Irfan A. Essa. Controlled animation of video sprites. In *ACM SIGGRAPH Symposium on Computer Animation*, pages 121–128, July 2002.

[3] Arno Schödl and Irfan A. Essa. Machine learning for video-based rendering. In *Advances in Neural Information Processing Systems. MIT Press*, volume 13, pages 1002–1008, July 2001.

[4] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics*, 22(3):277–286, July 2003.

[5] Okan Arikan and D. A. Forsyth. Synthesizing constrained motions from examples. *ACM Transactions on Graphics*, 21(3):483–490, July 2002.

[6] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, July 2002.

[7] Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics*, 21(3):491–500, July 2002.

[8] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: A two-level statistical model for character motion synthesis. *ACM Transactions on Graphics*, 21(3):465–472, July 2002.

[9] Ronald Metoyer. Building behaviors with examples. In *Ph.D. Dissertation, Georgia Institute of Technology*, 2002.

[10] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH 97*, Computer Graphics Proceedings, Annual Conference Series, pages 353–360, August 1997.

[11] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics*, 21(3):388–398, July 2002.

[12] D. A. Forsyth and J. Ponce. Computer vision: a modern approach. *Prentice-Hall*, 2001.

[13] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel. Finding motifs in time series. In *Proceedings of the Second Workshop on Temporal Data Mining*, pages 53–68, Edmonton, Alberta, Canada, July 2002.

[14] B.Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *ACM Conference on Knowledge Discovery and Data Mining, (KDD'03)*, pages 493–498, 2003.

[15] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, volume 26, pages 35–42, July 1992.

[16] H.-Y. Shum Z. Liu, C. Liu and Y. Yu. Pattern-based texture metamorphosis. In *Proceedings of Pacific Graphics 2002*, October 2002.

[17] Joel Carranza, Christian Theobalt, MArcus A. Magnor, and H. Seidel. Free-viewpoint video of human actors. *ACM Transactions on Graphics*, 22(3):569–577, 2003.

[18] T. L. Kunii, Y. Saito, and M. Shiine. A graphics compiler for a 3-dimensional captured image database and captured image reusability. pages 128–139.