Action Synopsis: Pose Selection and Illustration

Jackie Assa Yaron Caspi Daniel Cohen-Or
School of Computer Science Tel Aviv University

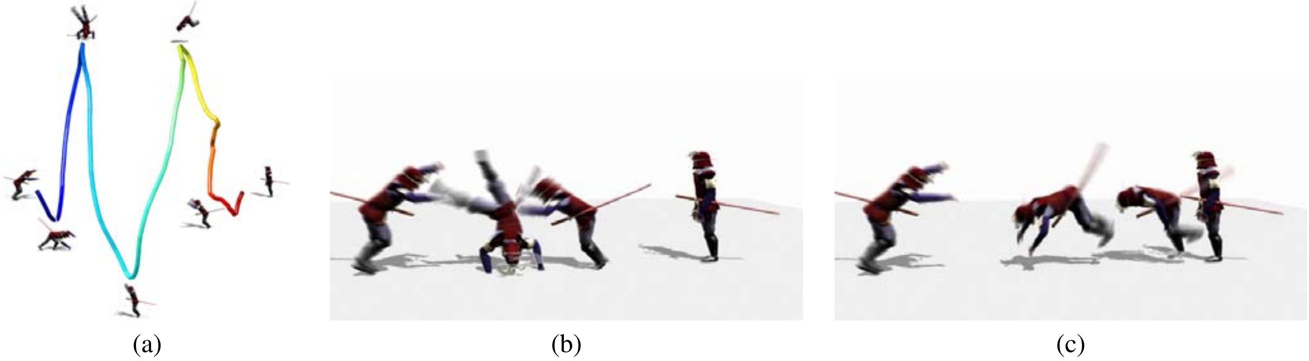


Figure 1: Action synopsis analyzes the motion-curve embedded in a low dimensional space (a); super-positioning of carefully selected poses (b) vs. uniform sampling (c). (Images used courtesy of Moshe Mahler and Jessica Hodgins, copyright Carnegie Mellon University.)

Abstract

Illustrating motion in still imagery for the purpose of summary, abstraction and motion description is important for a diverse spectrum of fields, ranging from arts to sciences. In this paper, we introduce a method that produces an action synopsis for presenting motion in still images. The method carefully selects key poses based on an analysis of a skeletal animation sequence, to facilitate expressing complex motions in a single image or a small number of concise views. Our approach is to embed the high-dimensional motion curve in a low-dimensional Euclidean space, where the main characteristics of the skeletal action are kept. The lower complexity of the embedded motion curve allows a simple iterative method which analyzes the curve and locates significant points, associated with the key poses of the original motion. We present methods for illustrating the selected poses in an image as a means to convey the action. We applied our methods to a variety of motions of human actions given either as 3D animation sequences or as video clips, and generated images that depict their synopsis.

CR Categories: I.3.7 [Computer Graphics]: Three-dimensional graphics and realism—Animation.; I.3.3 [Computer Graphics]: Picture/Image generation—Displaying and viewing algorithms.

Keywords: human motion analysis, animation analysis, key poses, dimensionality reduction, motion curve

1 Introduction

Expressing motion in still images, sculptures and monuments has intrigued artists and scientists alike for centuries. From Ancient Greece to the 19th century, human motion was captured mainly through instants of motion, which were presented by such techniques as interesting asymmetric poses, or by use of artistic elements such as hair and clothes swaying in the wind, which contributed to the sensation of movement [Ward 1979; Braun 1992]. Pioneers in the field of Photography also addressed the challenge of expressing human motion as an image. A technique known as Chronophotography captures a sequence of still images with a short time difference to allow displaying fast human activities. Stroboscopy [Cutting 2002] is currently used as a photography technique where the same film is used repeatedly at different time instances. The resulting image consists of a composition of frames with predefined time intervals which might not successfully convey the full motion, or perhaps cause self-occlusion.

The need to display human activity in still images exists in many fields of modern life. Illustrating a short sport event in a newspaper, or a dynamic experiment in a scientific journal, are examples of the need to display dynamic action in the printed media. Another example is a thumbnail of a video clip or animation stored in digital libraries. Typically, the first frame is used to represent the content. The method presented in this paper is a step toward improving the informative values of still images to convey short actions.

The method introduced in this paper, automatically selects poses that best represent an action for creating an *action synopsis*, allowing expression of complex motions in a still imagery. The main challenge in our work is to carefully select a small number of key poses taken from an animation sequence (3D) or a video clip (2D), and to present the associated frames in a concise view, in which the full motion is self explanatory. This is demonstrated in Figure 1, where the four key poses, selected by our technique, convey the action better than the frames sampled with uniform time intervals.

Studies by Kovar et al. [2002; 2004] Park and Shin [2004], and Lee et al. [2002] attempted selecting key-poses mainly for motion synthesis and retrieval, however the number of key-poses they ex-

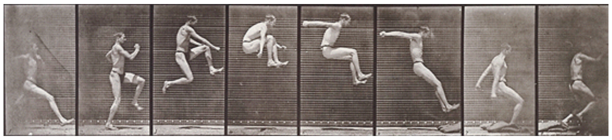


Figure 2: A locomotion series by Muybridge from the 19th century capturing human motion. Image courtesy of the Muybridge collection. Copyright of the National Museum of American History, Smithsonian Institution. http://www.si.edu/.

tract is more than an order of magnitude higher than can be used in an image composition such as ours. Similar work has attempted to select key-frames from video sequences based on image space criteria [DeMenthon et al. 1998; Vermaak et al. 2002; Fauvet et al. 2004]. Their approaches focus on the background scene and on the camera motion. In our work we handle a different media - animation sequences, focusing on a skeleton-based model activity.

Action synopsis selects key poses based on an analysis of the skeletal motion curve. The raw data that represents the skeleton motion is a high dimensional curve of poses, where each pose consists of the skeleton joints and their associated attributes or aspects. The key-idea is to successfully embed the high dimensional motion curve in a low dimensional Euclidean space that represents the actual motion well. The lower complexity of the embedded motion curve allows a rather simple geometric tool to analyze the curve in order to disclose significant points. The selected points along the curve are associated with the key poses of the original motion. The key-poses or the associated frames are used to synthesize an image or a small number of images for illustrating the action. The selection of the key-poses allows the generation of a clear understanding of the action, in particular, reducing self occlusions, adding means to express the action tempo, or selecting an optimal viewpoint.

The paper proceeds with a discussion of related work in Section 2, and a general overview of the proposed framework in Section 3. Sections 4 to 6 describe in detail the proposed method. Section 7 shows results and describes possible applications. We conclude in Section 8.

2 Background

With the dawn of the motion pictures era in the late 19th century, the study of high-speed optical phenomena began. At that time, photography entrepreneurs such as Muybridge, Marey, Duchamp and others started examining the behavior of movement by using a fixed intervals multiple-exposure technique named Chronophotography [Massey and Bender 1996; Cutting 2002; Ward 1979](see Figure 2 for a sample set of images presenting a male jump). This technique was refined with the introduction of stroboscopic photography by Edgerton [Kayafas and Jussim 2000]. Advances in digital photography allow synthesizing a multiple-exposure image from a video sequence. This provides the flexibility to select salient frames after the sequences have been captured, and not predefine them as in stroboscopic photography. It also makes it possible to compensate for camera motion and generate a wide field of view image known as "panorama" or "video mosaic" [Szeliski and Shum 1997]. These representations are shown in [Irani and Anandan 1998] to be useful for video indexing. The index is a composition of dynamic objects from each frame, superimposed on the static background panorama. To avoid an overlap between portions of the preceding foreground image objects, the video is diluted by overlooking many frames. This is usually done by sampling frames at a fixed rate, or manually [Massey and Bender 1996].

Selecting representative frames is not limited to video mosaics. Representative frames/poses, typically denoted as “key-poses” or “key-frames”, have been used in many domains of computer graphics (animation) and computer vision (video). Prominent applications include: activity detection and recognition, video or motion retrieval and motion synthesis or composition.

Activity detection, event recognition, and gait recognition usually uses key poses/frames to speed up the matching of a motion sample to a library of preprocessed examples. The representation of a single sequence varies from method to method, but many of them share the approach of clustering poses/frames and matching the new sample to the cluster representatives/prototypes. For example, Campbell et al. [1995] used phase-space (2D projections of joint positions and velocities) for recognizing atomic ballet moves, while [Zelnik-Manor and Irani 2001] used histograms of normalized space-time derivatives to detect predefined activities. Clustering is not the only approach for activity or activity style detection. In [BenAbdelkader et al. 2004], gait motion is represented by repeated blocks in an affinity matrix, constructed from distances between sequences of simultaneously scaled silhouettes. Elgammal and Lee [2004] used poses embedded in a low dimensional unit circle to compute key poses and an interpolation function to resample pose sequences. These temporally aligned sequences are than decomposed into a bilinear form that corresponds to style and content.

In a similar manner, using key frames and key poses is applied to video and motion summaries and retrieval. Cooper et al. [2002] decompose an affinity matrix in order to generate video summaries. Their affinity matrix is constructed from distances between DCT coefficients of video frames. The decomposition objective is to identify frames that on average are most similar to other frames. Vermaak et al. [2002] maximize the dissimilarity between consecutive key frames, and favor frames with high entropy.

Clustering as a means for selecting key frames has also been applied to skeletal motion. For example, Loy et al. [2003] select key frames that are centers of frame clusters. Their metric is based on contour matching and is used for retrieval of sport events video clips. Liu et al. [2003] store the extracted cluster representative key frames in an efficient motion index tree, to improve retrieval time of 3D motions with different speeds. Selecting cluster representatives does not comply with the objective of this work. As shown later, activity is better represented by local extremum points which tend to be off the center of the motion clusters.

Curve simplification has also been proposed for selecting key frames. Curve simplification [Ramer 1972] constructs a polygonal approximation to a curve, by repeatedly splitting the line at the point with the maximal distance from the curve. DeMenthon et al. [1998] applied curve simplification to video DCT coefficients, while [Lim and Thalmann 2001] applied it to 3D motion capture data. They reported that a reasonable approximation can be made with about a fifth of the frames. Note that in most clips we require a selection of less than 3% of the original poses. In such cases, it can be easily shown that the polygonal approximation may not select the best representatives as its distance from the original curve is not monotonically decreasing. Furthermore, both studies did not apply dimensionality reduction, which we believe is essential for this task, especially when the objective is choosing a small number of poses. Finally, Kunio and Matsuda [2004] looked at 3D motion data as a set of curves, and used handwriting techniques to detect key poses at every joint independently.

Key poses have also been used to speed up motion composition and synthesis. In this domain the objective is to find transition poses that can be blended to form a seamless transition between predefined motions. The result is typically represented by a "Motion Graph". Kovar et al. [2002; 2004] detected such seamless transitions candidates using distances between clouds of points driven by a skeleton, while [Grochow et al. 2004] propose to learn the distance function of a particular style. Lee et al. [2002] detected the transition locations between human motion sequences using trans-

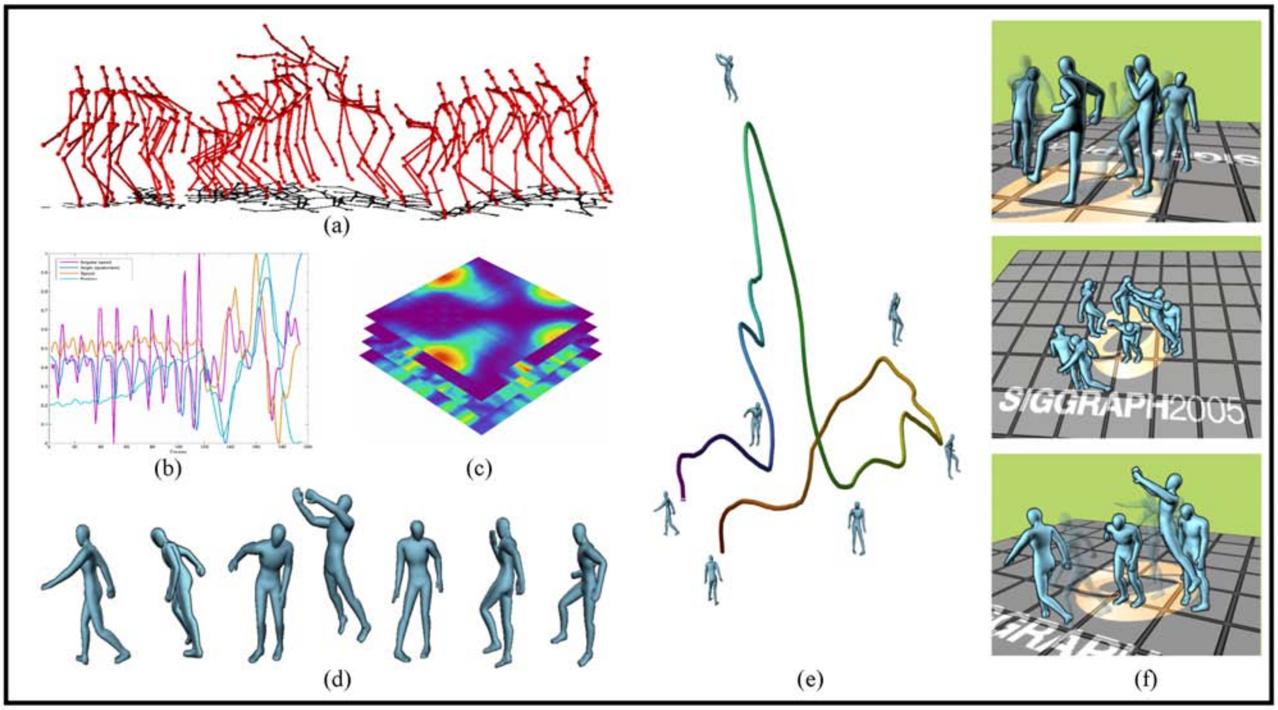


Figure 3: An overview of action synopsis: The input animation sequence (a) is analyzed. Aspects of a single joint (character's elbow) are displayed in (b) as a function of time. Affinity matrices (dissimilarities between poses) of the aspects are computed using all joint data (c). Key-poses (d) are extracted by embedding the motion curve in a low dimensional Euclidean space. Key poses are local extremum points of the motion curve (e). Synopsis views (f) display a superposition of key-poses.

lation and weighted average of rotation vectors of joint angles. Optimization of particular weights is reported in [Wang and Bodenheimer 2003]. Such distance measurement is similar to the one described in this paper, however we foresee the usage of other motion attributes as well.

Retargeting motion, is an additional goal utilizing the extraction of key frames. In several examples the task is performed in a low dimension space however the input data (i.e, joint positions) is first enriched with additional sets of motion measurement. Park and Shin [2004] used PCA coefficients computed from normalized motion capture data, and used cluster representatives to interpolate all other poses. They use in addition to the joint positions, also velocity, acceleration and dihedral angle. Safonova et al.[2004] reported that many human actions are intrinsically low dimensional. Using this observation they optimize human motion in a low dimensional space to synthesize realistic human motions. Similarly, here we also exploit the fact that motions are well presented with a relatively small number of dimensions.

3 Overview

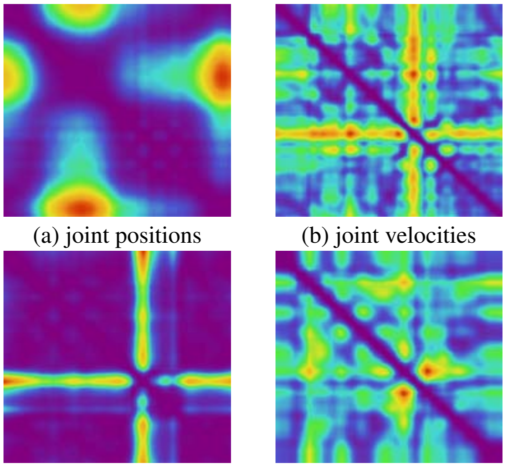
An animation sequence consists of a series of skeletal poses. The skeleton is defined by a number of skeleton parts connected by joints. With each joint we associate *aspects* which are important attributes derived from the given data. The challenge in the analysis is to identify adequate aspects, together with appropriate weights, so that it consists of vital motion information that can be revealed by the following steps. The animation sequence defines a high dimensional *motion curve*, where the dimension is defined by the number of skeleton parts and the number of aspects associated with each part. The high dimensionality is an obstacle here in the sense that applying the analysis directly to a high dimensional curve is, in some cases, ineffective as illustrated in Figure 11. The key-idea is to embed the curve in a low dimensional space in which a rather simple geometric algorithm can identify important poses. Hereafter, the term motion curve is used both to refer to the high dimensional and low dimensional motion curve, and we will refer to their dimensionality only whenever it won't be understood from the context.

A common approach for reducing the dimensionality is to express all the pose to pose distances in an affinity matrix and project the space spanned by the matrix columns into a small subspace spanned by the leading eigenvectors of the matrix. However, defining a distance measure between poses is not straightforward, nor is defining a single affinity matrix, since there are various "aspects" of a pose that might be important in different cases.

To alleviate this problem, we define a small number of affinity matrices, each describing a single aspect of the inter-pose distance (see Figure 3(c)). We empirically identify a small number of effective aspects that represent the motion well, where each aspect is assigned a weight which represents its relative importance. Points in a low dimensional space are produced from these affinity matrices by an RMDS method, which is a non-linear optimization process that reduces the dimensionality of multiple affinity matrices. Section 4 elaborates on the construction of the affinity matrices and the dimensionality reduction.

The low dimensional motion curve is then analyzed by identifying extremum points which are not close to each other. We use an iterative algorithm which selects points at local extrema. The selected points along the motion curve define the key-poses. The iterative algorithm defines a hierarchy where the top levels contain poses which are good representatives of the action.

The technique consists of the following main stages:
Extracting motion aspects. The motion data can be obtained from motion capture data, standard video clips, or readily available an-



(a) joint positions (b) joint velocities

(c) joint angles (d) joint angular velocities

Figure 4: The four affinity matrices. Each matrix captures the dissimilarity among the poses aspects. The different contribution of each aspect is evident from the differences in the location of most dissimilar poses.

imation sequences. We assume that the motion represents a short sequence of human action. In case of a video sequence, the skeleton of the participating human motions is tracked and extracted using standard tools [Gibson et al. 2003]. Our method uniformly treats 3D skeletons and 2D skeletons extracted from videos. Given a sequence of skeletal poses, a small number of motion aspects is calculated. A motion aspect is an attribute of the motion, by which we define inter-pose distances. In our implementation, we use four aspects, namely, joint positions, angles, speed and angular velocity. However, other aspects can be effective in various cases.

Dimensionality reduction. For each motion aspect we define an affinity matrix, which encapsulates weighted distances among all poses. The dimension of the motion curve is reduced by analyzing the affinity matrices with a Replicated Multi-dimensional Scaling (RMDS) [McGee 1978]. The technique defines a reduced dimensional space (typically of 5-9 dimensions) in which the salient features of the various motion aspects are kept.

Pose selection. In this stage we locate the local extreme points along the motion curve, which are associated with the extreme poses of the motion. This stage generates a hierarchy of prioritized poses.

Synopsis illustration. In the last stage, the frames associated with the selected poses are composed into an image. The selected frames can either be presented side by side, or be composed into a single image. To further enhance the image, some instances can be rendered semi-transparently, thus reducing the cluttering of the resulting image and highlighting the more significant poses. The following sections provide a detailed description of each step.

4 Motion Curve

4.1 Extracting motion aspects

An input motion sequence is a series of joint positions of a single skeletal object. The positions may be extracted from 2D (video) or 3D (animation). We refer to each set of joints in a given frame as a *pose*. From this data we generate additional attributes of the skeletal motion, which together form the motion aspects. In our implementation they include: (i) joint positions, (ii) joint angles, (iii) joint velocities, and (iv) joint angular velocities. Similar motion aspects are used by Grochow et al.[2004] for determining motion similarity. The analysis of different types of motion may require including other aspects. However, in our experiments the above four aspects were found to be sufficient, whereas some other aspects such as acceleration are found to be ineffective for our purposes.

The joint positions (about 18-25 joints) are derived from the raw data. The joint 3D angles (about 8-12 skeletal joints which have two skeleton neighbors) are measured using rotations of unit quaternion. Joint velocity is approximated using the position differences between the pose before and after the given frame. Joint angular speed is approximated using Euclidian distance between the two unit quaternion. To smooth motion capture data, we used the Loess method [Cleveland and Devlin 1988], also known as locally weighted polynomial regression. Figure 3(b) depicts the four aspects of the joint of the right elbow in the 'walk-jump' sequence illustrated in the figure. As discussed below, different aspects contain diverse vital information for the analysis of the importance of the various poses.

4.2 Affinity matrices

To deal with the inherent high dimensionality of the given motion data, we employ a dimensionality reduction. We first describe this step, and then discuss its advantages. For each motion aspect we generate a separate affinity matrix. Let $\mathbf{x_a^f}$ be the vector representing a aspect of the pose at frame f. The dissimilarity of this aspect (d_a) between two given frames f_1 and f_2 is computed by a weighted standardized Euclidean distance:

$$d_a(f_1, f_2) = \sum_{j \in joints} b_j \frac{(x_j^{f_1} - x_j^{f_2})^2}{\sigma_j^2}, \tag{1}$$

where each coordinate in the sum of squares is inversely weighted by σ_j^2 the variance of that joint coordinate. For the sake of clarity, the index a is omitted from the right hand side. Alsothe contribution of different joints to the overall dissimilarity is weighted by b_j according to their respective importance. These weights are derived from the skeletal joint hierarchy, as well as from the limb importance.

In some cases when the dimensionality reduction output does not approximate the high dimension distances (see Section 4.3) we enhance the significance of the temporally close poses by introducing a *time window*. This is realized by multiplying the original dissimilarity of Eq. 1 ($d(f_1, f_2)$) by an exponential decay function, which reduces the weights of the temporally distant frames:

$$\hat{d}(f_1, f_2) = e^{-\frac{(|f_1 - f_2|)}{N}} d(f_1, f_2), \tag{2}$$

were N is the sequence length. Figure 4 displays the four affinity matrices of the 'walk-jump' sequence shown in Figure 3. The activity in this sequence includes a walk with an embedded jump and hop actions. The values of the affinity matrices are expressed as colors, where hot colors represent a large dissimilarity. It is rather apparent that the matrices represent different pose dissimilarity behavior and consequently different key poses are emphasized. Specifically, the position aspect in (a) presents the distances in position between poses. The walking in the sequence is along a "horse shoe" path, and the hot region corresponds to the two poses positioned most farther apart. Velocity in (b) highlights the difference in speed between the walk (slow) and the jump (fast) motion, whereas the joint angles aspect (c) highlights the difference between the upper point of the jump, in which the skeleton is stretched and the rest of the walking cycle, in which the skeleton is more relaxed. Finally, the angular velocity in (d) prefers the preparation for the jump, preparation crouch and the landing postures.

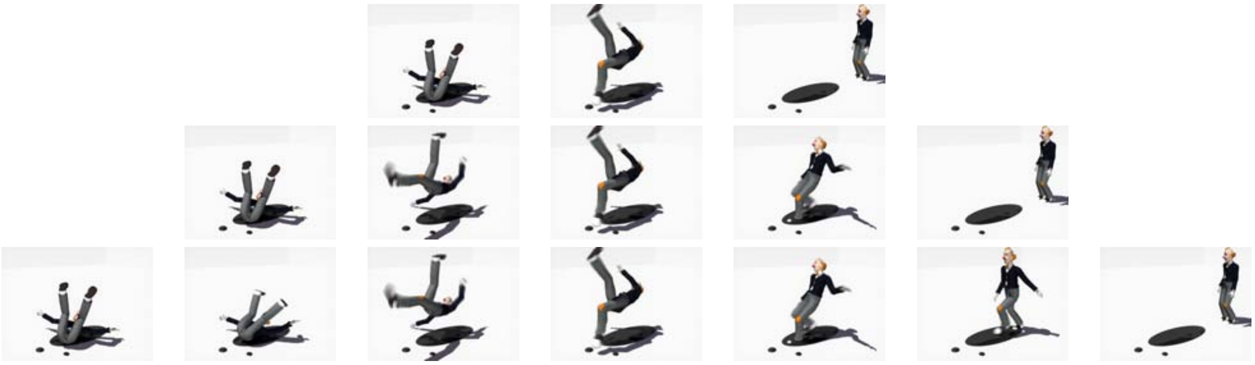


Figure 5: The iterative selection of points leads to a hierarchical set of key-poses. The three rows show a selection of 3,5 and 7 poses respectively. The motion curve for this sequence is shown in Figure 6. (Images used courtesy of Moshe Mahler and Jessica Hodgins, copyright Carnegie Mellon University.)

4.3 Replicated multi dimensional scaling

To reduce the number of dimensions we use non metric replicated multi-dimensional scaling (non metric RMDS), an extension of multi-dimensional scaling (MDS), which originated in the mid 20th century [Young and Householder 1941; Torgerson 1952]. Using a dissimilarity matrix between elements in high dimensional space X, generates an embedding into a low dimensional Euclidean subspace $\phi : X \rightarrow R^d$. Denote by $d_{i,j} = d(x_i, x_j)$ the dissimilarity measure in X, and by $e_{i,j} = ||\phi(x_i) - \phi(x_j)||$ the Euclidian distance in the low dimension Euclidian space. The objective of MDS is to define a mapping ϕ that best expresses the distances $d_{i,j}$ with $e_{i,j}$. That is:

$$\min_{\phi} \sum_{i,j} (e_{i,j} - d_{i,j})^2 \tag{3}$$

Shepard [1962] and Kruskal [1966] extended MDS to allow handling of non-metric data. The similarity between distances in high and low dimension are preserved only up to a monotone increasing function f. That is, non-metric MDS attempts to minimize the following:

$$\min_{\phi} \sum_{i,j} w_{i,j}(f(e_{i,j}) - d_{ij})^2, \tag{4}$$

where f is a monotone scalar function. RMDS [McGee 1978] allows us to analyze multiple affinity matrices by finding a mapping ϕ that optimally approximates the distances of multiple aspects simultaneously. Because it uses a different monotone increasing function for each matrix, it implicitly introduces a relative weight for each initial dissimilarity measure.

An important topic in dimensionality reduction techniques is determining the desired low dimensional goal. Kruskal [1966] proposes to measure the fitness of the output by calculating its stress, namely, considering the dissimilarities between the distances between the low dimension points, to all k initial affinity matrices:

$$\sum_{k} \sum_{\mathrm{i<j}} w^k_{ij}(f^k(e_{ij})) - d^k_{ij})^2. \tag{5}$$

Analyzing the stress as a function of the dimension, helps define the target number of dimensions. In most cases, a reduction to 5-8 dimensions provides pleasing results, which comply with the results reported in [Safonova et al. 2004].

5 Pose Selection

The output of the RMDS algorithm is the low dimension motion curve denoted here by $C(p)$. It is comprised of the low dimension points ordered according to their original temporal order. This section describes the selection of the key-poses along $C(p)$. We argue that high local extremum points are the best representative poses. In their selection, we consider the distance from their neighborhood (measure of local extremum), and their temporal distance to other selected key poses. The method allows selecting a prescribed number of points or defining their number according to a predefined error threshold.

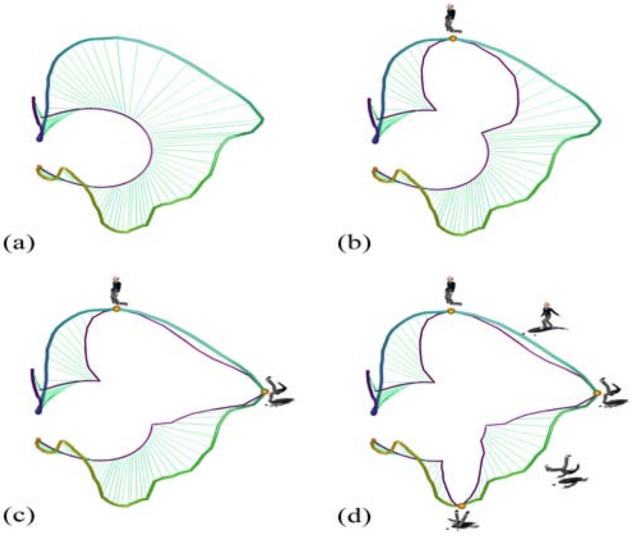


Figure 6: Steps of the pose selection algorithm. (a) The initial motion curve and the average curve (in purple); the distances to the averaged curve are indicated by the straight lines. (b-d) shows the selection of the highest distance points, and the modification of the average curve with their respective selected poses.

A point in $C(p)$ is projected into $\bar{C}(p)$ - the weighted average location of its neighboring points:

$$\bar{C}(p) = \sum_{i \in \delta} C(i) e^{-\frac{\|(p-i)\|^2}{\delta^2}} / \sum_{i \in \delta} e^{-\frac{\|(p-i)\|^2}{\delta^2}}, \tag{6}$$

where δ is a small constant window around $C(p)$. The sequence of all the points in $\bar{C}(p)$ forms a smooth version of the curve $C(p)$. Denote by r_p the above distance:

$$r_p = ||C(p) - \bar{C}(p)||. \tag{7}$$

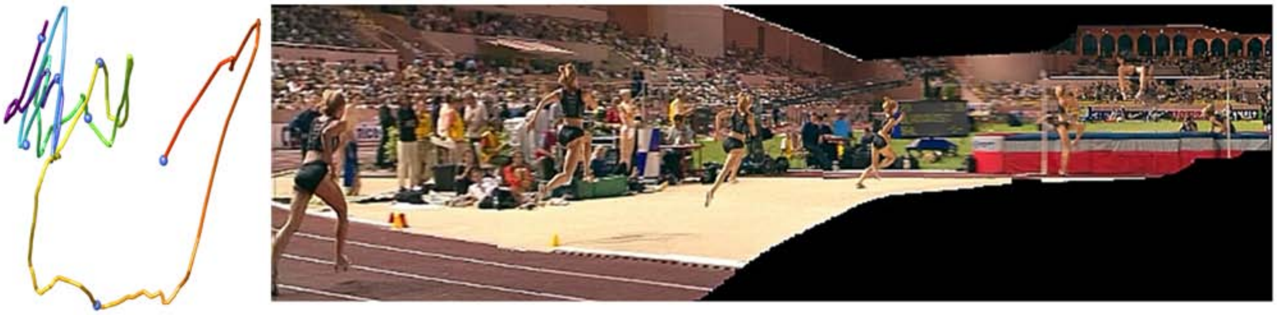


Figure 7: Panoramic views of a height jump. The images are stitched using graph cuts on a common panoramic background generated from a video sequence (see text). Images used courtesy of Israel Sport5 channel.

The two curves and the distances between them are illustrated in Figure 6. The algorithm selects a set of points with large values of r_p such that they are temporally distant apart. The following iterative steps are performed until sufficient frames are selected, or until the maximal r_p is below a predefined threshold (i) Select the point p_i with the largest r, and add it to the to the list of selected points; (ii) Modify $\bar{C}(p)$:

$$\bar{C}(p) := \alpha C(p) + (1-\alpha)\bar{C}(p), \tag{8}$$

where $\alpha = e^{-\frac{\|i-p_i\|^2}{\gamma^2}}$, and γ is a small constant time window. Usually the values of δ are determined by the ratio of the arc lengths of $C(p)$ and $\bar{C}(p)$. This ratio indicates whether the $\bar{C}(p)$ sufficiently averages the original curve $C(p)$. The value of γ is determined by the number of poses, the desired number of selections and the polling resolution. Increasing the number of selection points decreases the window size up to a certain threshold depending on the polling resolution. The algorithm is illustrated in Figure 6. The curve $\bar{C}$ is in purple and the figure visualizes the distances between the points in C and their projections on $\bar{C}$. The output of the above process, is a list of key poses ordered by their importance. It may also be viewed as a non-balanced binary partitioning tree, where the nodes represent intervals between key poses. Each interval can be split, and hence produces two sub nodes. Such an output is illustrated in Figure 5. One important attribute of our selection method is that it allows adding priors for the selection of one or more poses at any desired level. The algorithm adapts to the prior by selecting the initial preselected set of points and adapting the smooth curve accordingly. This allows further selections to account for the preselected points in a natural manner. Additional constraints can be placed by adding weights to the initial smooth curve according to different segments of the sequence, to allow improving the selection from existing segmentation methods.

6 Synopsis Illustration

Given a set of key-poses possibly in the form of a hierarchy, we now describe methods for illustrating the action synopsis in an image or a set of images. Maybe the simplest way is to display the views side by side. An example is shown in Figure 5. This type of display is used in comics or in storyboards. Having the key-poses in 3D allows us to generate views of key-poses by rendering a virtual character from any point of view. To combine them into an image, one has to take into account view-dependent factors such as the scene background illustrating the context of the activity, self occlusion, or the extent of the image dimensions.

To reduce image space redundancy, a number of instances can be combined sharing a common background. Figure 1 shows an example of a digital strobing, where views of four key-poses are combined into the image. When these views are rendered, the camera should be placed in such a way that the instances are spread as wide as possible. This requires us to identify the principal motion axis and place the camera in perpendicular direction to that axis.

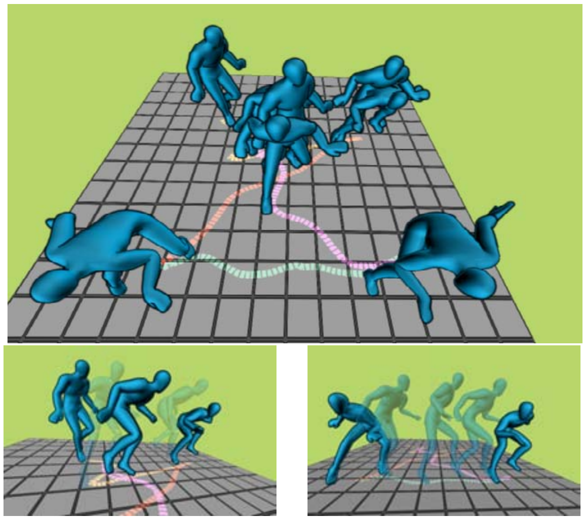


Figure 8: The sneaking sequence as an example of a complex action which requires segmentation into several views.

Our algorithm for digital strobing identifies image space redundancy and adds more views if possible. The most important poses are composed first. Poses that are selected later are add with high transparency, such that they will not occlude the most important key poses (see Figures 8 and 13). We use the simple image silhouette to measure the amount of poses overlapping among the instances to avoid self-occlusion.

If the overlapping is large, the instance can be rendered semi-transparent as shown in Figure 8. Note that digital strobing is prone to temporal ambiguities. A man jumping backwards is likely to be perceived as jumping forward. Adding an arrow or various symbolic means can alleviate such ambiguities. Similarly, when the action is long and too complex, more than one image is required to understand the full motion. Figure 8 shows three images of a sneaking sequence. In that sequence the character is moving around with a non-trivial trajectory. Such a long motion requires applying a segmentation method and displaying the key poses from each segment separately. The complex trajectory also requires to render some auxiliary markers to assist its understanding and the spatial relation among the different views.

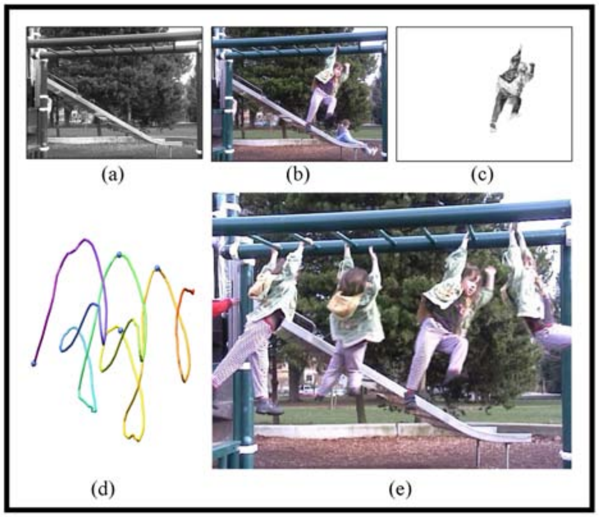


Figure 9: Four selected key poses from the monkey bar sequence are superpositioned to illustrate the activity. One of them is shown in (b) with its distance from the temporal median image (a) shown in (c). The sequence motion curve is presented in (d), and the resulting composition (e). The sequence is courtesy of Michael Cohen and his daughter Lena.

6.1 Video sequences

To illustrate action synopsis of video sequences, it is required to extract the moving foreground objects from the selected frames, and stitch them into a single image. For simple cases, when the camera and background are static, this can be done implicitly. A background image is generated by taking a pixel's median value over all the frames of sequence. Figure 9(a) is an example of a median image. The selected key frames can then be blended in, using weights proportional to the distance (in RGB space) between pixel values in each frame and the median image (e.g., Figure 9).

When the camera is moving, we first align all frames into a common coordinate system, using the method described in [Szeliski and Shum 1997], and then compute a median image. Similar to [Agarwala et al. 2004] our technique uses graph cuts to stitch parts from selected key poses/frames into the background panorama (the sequence median).

The method automatically marks bounding blobs around moving objects, and applies a graph cut to paste them into the background panorama. For every selected key frame I_k we threshold the image defined by the multiplication of forward and backward difference images: $(I_{k-t} - I_k)(I_{k+t} - I_k)$, apply morphological cleaning, and connected-component analysis to identify the bounding blob, that is a region of motion pixels that must be included in the output image. The background panorama that is not visible in this frame is also a part of the output. Finally, graph cut is used to select an optimal seam between the blob and the current frame. An example of this graph cut stroboscopic mosaic is shown in Figure 7.

Image mosaics are hard to construct for activities which spread along a wide area in the scene, when the scene contains substantial depth variations (parallax). For example, placing all key poses of the pole vault sequences (see Figure 12), yields an inefficient use of space and large distortions. Note that the long run prior to the jump requires that a spatially large area without many key poses be captured.

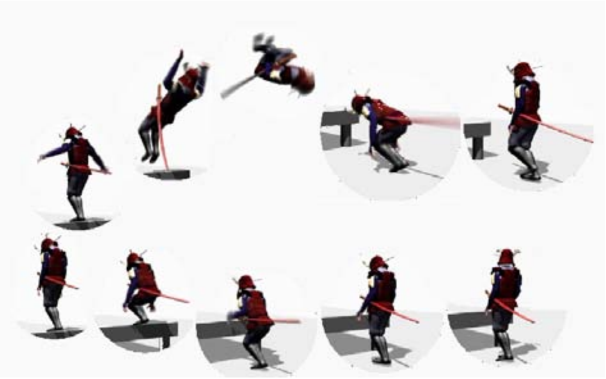


Figure 10: A back flip sequence illustrated with a digital stroboscopic spatially expanded layout. The technique allows to express spatially large or small and temporal sequences into a concise view. (Images used courtesy of Moshe Mahler and Jessica Hodgins, copyright Carnegie Mellon University.)

6.2 Spatially expanded layout

Digital strobing and image mosaics fail to produce pleasing results when different poses occlude each other. We address such cases by placing key poses in a layout that does not preserve their original absolute spatial location, but which preserves the overall structure of the activity. This is denoted here as "spatially expanded layout" and is illustrated in Figure 10.

The method preserves the local orientation among the pose locations in the foreground images so that any movement among them is still being perceived in the composition. For each selected frame we define an elliptical 'area of interest', which bounds the instance in action. Their placement in the images is determined by applying a simulated annealing algorithm designed to minimize the energy of a given configuration. The energy is comprised of of three terms: (i) Consecutive poses orientation. This term attempts to keep the orientation angle between consecutive poses. (ii) Non consecutive poses distance. This force acts among any non-consecutive but spatially adjacent frames, and decays after a certain distance. This term expands the resulting configuration. (iii) Consecutive poses distance. This force pulls two consecutive poses to a predefined proximity range. In the same manner, consecutive poses which occlude each other are being pushed apart.

Figure 10 displays stretched views from a video sequence of a forward jump on a table and a back flip. Without the stretching, the views will occlude each other. The location of each pose was computed by minimizing the above energy function terms.

7 Results and Applications

7.1 Implementation details

We have implemented the action synopsis method and tested it over a large amount of motion data, most of which are 3D motion capture data. For our tests we use available motion capture data repositories such as Carnegie Mellon University - CMU Graphics Lab - motion capture library, stockmoves by Motek BV, and bvhfiles by Animazoo. For video clips we extracted the 2D joints data with Icarus software, which provides a semi-automatic tracking of features. Although tracking might yield a non-accurate data, the key benefits of using action synopsis is that it does not require highly accurate skeleton extraction, and that it tolerates joint occlusion. The dimensionality reduction is performed using the PROXSCAL implementation provided by SPSS software.

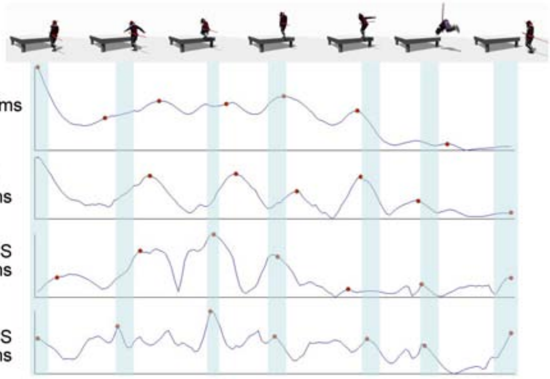


Figure 11: Difference in selections between several motion curve alternatives. The graphs display the distance between motion curve and its averaged curve (r_p in Eq. 7) without dimensionality reduction, and with 3 and 9 dimensions applying either PCA or RMDS methods. Note the difference in extreme points, and in the selected key frames (marked in red) or the alternatives. The cyan bars show the "ground truth" selection (described in Section 7) with their corresponding key poses on top.

The data pre-processing and post-processing including the pose selection algorithm is implemented using Matlab. As a final step the corresponding frames are rendered using Kaydara Motion Builder, Curious Labs Poser and our own software. Several of the used rendered images were downloaded from the CMU Graphics Lab site.

7.2 Results and discussion

The examples studied using the action synopsis, included diverse types of animations and videos, among them: Simple actions - such as the 'walk-jump' sequences of Figures 3 and 5, intermediate complexity motions containing compound motions or longer sequences such as the ones showed in Figures 1 and 10, or complex motions as presented in Figure 8. The analysis of video sequences included both home video - Figure 9 and various sports fields such as pole vault and high jump shown in Figures 7 and 12

To evaluate the key frame selections made by various methods we have conducted a user study. Several motion video clips, were shown to a group of 25 adults with no prior experience working professionally with animation. Participants in the study were asked to select a specific number of frames best representing the complete sequence. For each clip, the results were compared with the selections made by our algorithm using the original data, using data reduced to 3 dimensions by PCA or RMDS, and data reduced to 9 dimensions by RMDS. The results for one of the given clips is shown in Figure 11. It shows the distance (r_p in Eq. 7) between the resulting motion curve and the averaged smooth curve (see Section 5). The red dots present the frames selected by our selection algorithm on each of the curves, whereas the cyan bars present the user selections along their standard deviation. More results are shown in Table 1.

Note that RMDS with 9 dimensions produces the best correlation to the users selection. It should be noted that in other clips with more complex motion, users selection were more spread throughout the clip. This was emphasized in cases such as the one presented in the monkey-bar sequence (Figure 9). The repetitive nature of the action is clearly visible in the motion curve. Thus, in this case, any sparse selection of poses can yield plausible synopsis.

A naive linear dimensionality reduction using a set of predefined weights, can be obtained by performing a PCA analysis on a linear combination of the motion data. This yields in a low dimensional motion curve which lacks some of the properties of the original motion, and thus is not as informative as our method. In our experiments RMDS also subsumed locally linear embedding (LLE). We believe that this is because LLE preserves the local geometry and does not provide a global analysis necessary for selecting the key-frames. Overall, RMDS provides us with additional attributes that are valuable in this problem domain. It allows overcoming missing data, to reduce the importance of temporally far frames, and to tune the contribution of various aspects.

Title	STD of selections	Number of automatically selected pose within one STD		
		87 Dims	PCA	RMDS
Backflip	3	3/7	3/7	7/7
Cartwheel	2	4/4	4/4	3/4
Walk-jump	5	3/7	3/7	7/7
Dance	22	N/A	N/A	N/A
Pole vault	12	3/4	3/4	3/4
Long jump	8	2/3	2/3	2/3
Height jump	5	4/6	3/6	5/7

Table 1: Comparison of different approaches to user study selection. It displays for each clip the average STD (in frames) of users key frames selection. The smaller this value the larger the consensus among users is. Thus, we did not report the results of the dance sequence (the ground truth is not reliable). The right 3 rows reports the number of key poses that each method selected that falls within one STD from the users mean selection.

7.3 Applications

The synopsis image can immediately be used as an icon or as a thumbnail of an animation sequence. Two examples are shown in Figure 13: (i) A stroboscopy illustration, and (ii) A single key pose. Note, that even using the key-pose that is selected first, is more informative than using the first frame (a standing person as shown on the right of the stroboscopy illustration). The concept from the user view point is illustrated in Figure 14, namely a view of a directory containing a few animation sequences. Action synopsis can be used to automatically and semi-automatically generate comic strips and story boards of existing animation and video sequences, by utilizing existing segmentation methods, and extracting the key poses from each segment, presenting the results in frame sequence or composition. An example for such a usage is shown in a short strip in Figure 15.

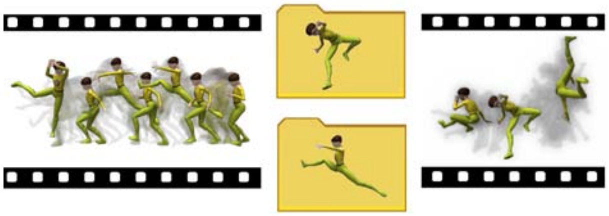


Figure 13: Thumbnails of a ballet and breakdance animation sequences. The activity is illustrated using a superposition of selected poses with different contrast. Low priority poses are rendered as ghosts (low contrast). Alternatively, only the first selected key pose is displayed.

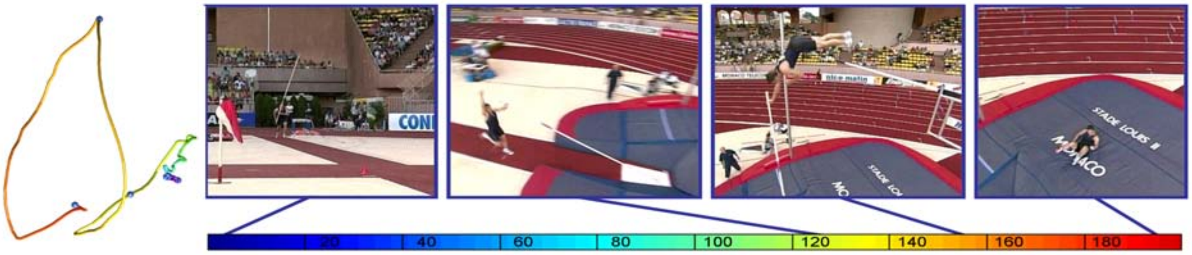


Figure 12: Four selected frames from a pole vault sequence. Note the the selected frames are all concentrated in the second half of the sequence. Images used courtesy of Israel Sport5 channel.

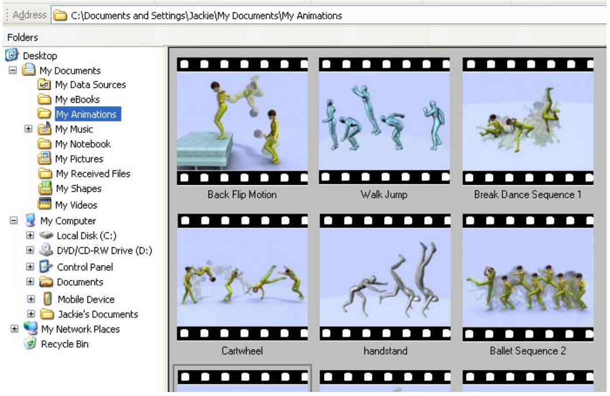


Figure 14: An animation directory illustrated using thumbnails of animation sequences.

8 Summary and Future Work

In this paper we focus on generating an action synopsis of a single skeleton-based character, using a few motion aspects. The resulting curve characteristics remain compatible with the initial motion characteristics in the sense that extreme model poses are being projected to extremum points in the low dimensional motion curve.

We focused on four useful aspects of the motion, but in essence, the ideas presented here can be extended to deal with other aspects of skeletal motion. Furthermore, they may be extended to activities/motions that are not necessarily provided as a sequence of skeleton poses. Technically, it suffices to define a similarity/dissimilarity measure (not necessarily metric) between consecutive frames. Then it would be possible to apply the action synopsis approach presented in this paper.

In the future we plan to analyze more complex motions which contain more than a single skeleton, such as team games, pair sports, etc. The frame selection algorithm can be extended to include more advanced selection rules, which, for example, also consider the relation between frames, prohibiting the selection of frames with occluded foreground images, or using background image difference as well. These rules can be represented by additional aspects, influencing the pose selection and the selection of frames that can be easily incorporated into a mosaic composition.

9 Acknowledgements

Part of the data and several of the motions rendering were obtained from mocap.cs.cmu.edu database which was created with funding from NSF EIA-0196217. This work was supported in part by grants from the Israel Science Foundation (founded by the Israel Academy of Sciences and Humanities), by the Israeli Ministry of Science.

References

AGARWALA, A., DONTCHEVA, M., AGRAWALA, M., DRUCKER, S., COLBURN, A., CURLESS, B., SALESIN, D., AND COHEN, M. 2004. Interactive digital photomontage. In *ACM Transactions on Graphics, (SIGGRAPH)*, 294–302.

BENABDELKADER, C., CUTLER, R., AND DAVIS, L. 2004. Gait recognition using image self-similarity. *EURASIP Journal on Applied Signal Processing 15*, 4, 572–585.

BRAUN, M. 1992. *Picturing Time*. U. of Chicago, Reading, MA.

CAMPBELL, L. W., AND BOBICK, A. F. 1995. Recognition of human body motion using phase space constraints. In *International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, 624–630.

CLEVELAND, W., AND DEVLIN, S. 1988. Locally weighted regression: An approach to regression analysis by local fitting. In *J. of the American Statistical Association*, vol. 83, 596–610.

COOPER, M., AND FOOTE, J. 2002. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*.

CUTTING, J. E. 2002. Representing motion in a static image: constraints and parallels in art, science, and popular culture. *Perception 31*, 1165–1193.

DEMENTHON, D., KOBLA, V., AND DOERMANN, D. 1998. Video summarization by curve simplification. In *Proceedings of the sixth ACM international conference on Multimedia*, ACM Press, 211–218.

ELGAMMAL, A., AND LEE, C. 2004. Gait style and gait content: Bilinear model for gait recognition using gait re-sampling. In *6th International Conference on Automatic Face and Gesture Recognition*, 624–630.

FAUVET, B., BOUTHEMY, P., GROS, P., AND SPINDLER, F. 2004. A geometrical key-frame selection method exploiting dominant motion estimation in video. In *Int. Conf. on Image and Video Retrieval, CIVR 2004*, vol. 3115 of *Lecture Notes in Computer Science*, 419–427.

GIBSON, S., HUBBOLD, R. J., COOK, J., AND HOWARD, T. 2003. Interactive reconstruction of virtual environments from video sequences. *Computers & Graphics 27*, 2, 293–301.

GROCHOW, K., MARTIN, S., HERTZMANN, A., AND POPOVIC, Z. 2004. Style-based inverse kinematics. In *ACM Transactions on Graphics, (SIGGRAPH)*, 522–531.

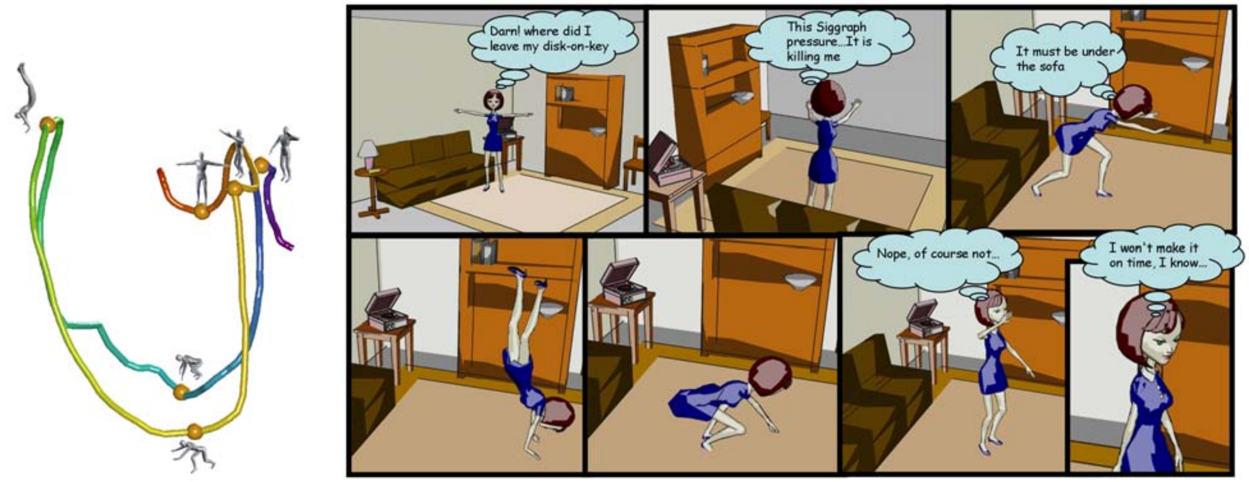

Darn! where did I leave my disk-on-key
This Siggraph pressure...It is killing me
It must be under the sofa
Nope, of course not...
I won't make it on time, I know...

Figure 15: "A day before the deadline". The selected key poses are rendered into a comic story board.

IRANI, M., AND ANANDAN, P. 1998. Video indexing based on mosaic representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence 86*, 5, 905–921.

KAYAFAS, G., AND JUSSIM, E. 2000. *Stopping Time : The Photographs of Harold Edgerton*. Harry N Abrams, New York, NY.

KONDO, K., AND MATSUDA, K. 2004. Keyframes extraction method for motion capture data. *Journal for Geometry and Graphics 08*, 081–090.

KOVAR, L., AND GLEICHER, M. 2004. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph. 23*, 3, 559–568.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002. Motion graphs. *ACM Transactions on Graphics, (SIGGRAPH) 21*, 3 (July), 473–482.

KRUSKAL, J. 1966. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1–27.

LEE, J., CHAI, J., REITSMA, P., HODGINS, J. K., AND POLLARD", N. 2002. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics, (SIGGRAPH) 21*, 3 (July), 491–500.

LIM, I. S., AND THALMANN, D. 2001. Key-posture extraction out of human motion data by curve simplification. In *23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, 1167 – 1169.

LIU, F., ZHUANG, Y., WU, F., AND PAN, Y. 2003. 3d motion retrieval with motion index tree. *Comput. Vis. Image Underst. 92*, 2-3, 265–284.

LOY, G., SULLIVAN, J., AND CARLSSON, S. 2003. Pose-based clustering in action sequences. In *Workshop on Higher-Level Knowledge in 3D Modeling & Motion Analysis*, 66– 72.

MASSEY, M., AND BENDER, W. 1996. Salient stills: Process and practice. *IBM Systems Journal 35*, 3/4, 557–573.

MCGEE, V. C. 1978. Multidimensionnal scaling of n sets of similarity measures : A nonmetric individual differences approach. *Multivariate Behaviour Research 3*, 233–248.

PARK, M. J., AND SHIN, S. Y. 2004. Example-based motion cloning. *Computer Animation and Virtual Worlds 15*, 3-4, 245–257.

RAMER, U. 1972. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing 1*, 3 (Nov.), 244–256.

SAFONOVA, A., HODGINS, J. K., AND POLLARD, N. S. 2004. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics, (SIGGRAPH) 23*, 3, 514–521.

SHEPARD, R. 1962. Analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika 27*, 125–139.

SZELISKI, R., AND SHUM, H.-Y. 1997. Creating full view panoramic image mosaics and environment maps. In *ACM Transactions on Graphics, (SIGGRAPH)*, 251–258.

TORGERSON, W. 1952. Multidimensional scaling: 1. theory and method. *Psychometrika 17*, 401–419.

VERMAAK, J., PIREZ, P., GANGNET, M., AND BLAKE, A. 2002. Rapid summarisation and browsing of video sequences. In *British Machine Vision Conference, BMVC*, vol. 1.

WANG, J., AND BODENHEIMER, B. 2003. An evaluation of a cost metric for selecting transitions between motion segments. In *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 232–238.

WARD, J. L. 1979. *Perception and Pictorial Representation*, vol. 1 of *The Art of Computer Programming*. Praeger, New York, ch. A piece of the action: Moving figures in still pictures, 246–271.

YOUNG, G., AND HOUSEHOLDER, A. S. 1941. A note on multidimensional psycho-physical analysis. *Psychometrika*, 331–333.

ZELNIK-MANOR, L., AND IRANI, M. 2001. Event-based analysis of video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 123–130.