Eurographics/ACM SIGGRAPH Symposium on Computer Animation (2004)
R. Boulic, D. K. Pai (Editors)

Animal gaits from video

Laurent Favreau, Lionel Reveret, Christine Depraz, Marie-Paule Cani

GRAVIR-INRIA

Abstract

We present a method for animating 3D models of animals from existing live video sequences such as wild life documentaries. Videos are first segmented into binary images on which Principal Component Analysis (PCA) is applied. The time-varying coordinates of the images in the PCA space are then used to generate 3D animation. This is done through interpolation with Radial Basis Functions (RBF) of 3D pose examples associated with a small set of key-images extracted from the video. In addition to this processing pipeline, our main contributions are: an automatic method for selecting the best set of key-images for which the designer will need to provide 3D pose examples. This method saves user time and effort since there is no more need for manual selection within the video and then trials and errors in the choice of key-images and 3D pose examples. As another contribution, we propose a simple algorithm based on PCA images to resolve 3D pose prediction ambiguities. These ambiguities are inherent to many animal gaits when only monocular view is available.

The method is first evaluated on sequences of synthetic images of animal gaits, for which full 3D data is available. We achieve a good quality reconstruction of the input 3D motion from a single video sequence of its 2D rendering. We then illustrate the method by reconstructing animal gaits from live video of wild life documentaries.

***Key words**: Animation from Motion/Video Data, Interpolation Keyframing, Intuitive Interfaces for Animation.*

1. Introduction

Traditional motion capture methods - either optical or magnetic - require some cooperation from the subject. The subject must wear markers, move in a reduced space, and sometimes has to stay on a treadmill. The range of possible captured motions is thus very limited: capturing the high speed run of a wild animal, such as a cheetah running after his pray is totally untractable using this method. This is unfortunate since this kind of motion data would be of great interest for 3D feature films and special effects, for which fantastic animals must be animated while no source of motion is available.

The new method we propose allows the extraction of 3D cyclic motion of animals from arbitrary video sequences (we are currently using live sequences from wild life animal documentaries). State of the art techniques in computer vision for markers-less 3D motion tracking are still hard to use in an animation production framework. As an alternative, we propose to use a robust existing techniques in a novel pipeline: we combine PCA of images and animation by interpolation of examples to reliably generate 3D animation of animal gaits from video data. PCA of images is well suited for animal gaits since this motion is naturally cyclic and PCA will factorize similar patterns and isolate main variation in images. Our experiments show what constraints and additional processing can be used to help PCA to focus on coding variation due to motion only. Our goal is to isolate and characterize, using PCA images, minimal sets of cyclic motion and to subsequently generate the associated 3D animation. More complex 3D animation with non uniformly cyclic motion could later be generated using recent methods in motion synthesis. We improve existing techniques with 2 main contributions: an automatic criterion to select examples from video and an algorithm to resolve ambiguities in the prediction of 3D poses from 2D video.

The resulting method greatly saves effort for the animator. Traditionally for the animation of quadrupeds, the artist must make several trails to set the key-frames and 3D poses. Our method, based on PCA images, allows us to provide directly the visually salient key-images with which to associate a 3D pose. The interpolation methods automatically generates long sequence of 3D animation mimicking the rhythm of the original video.

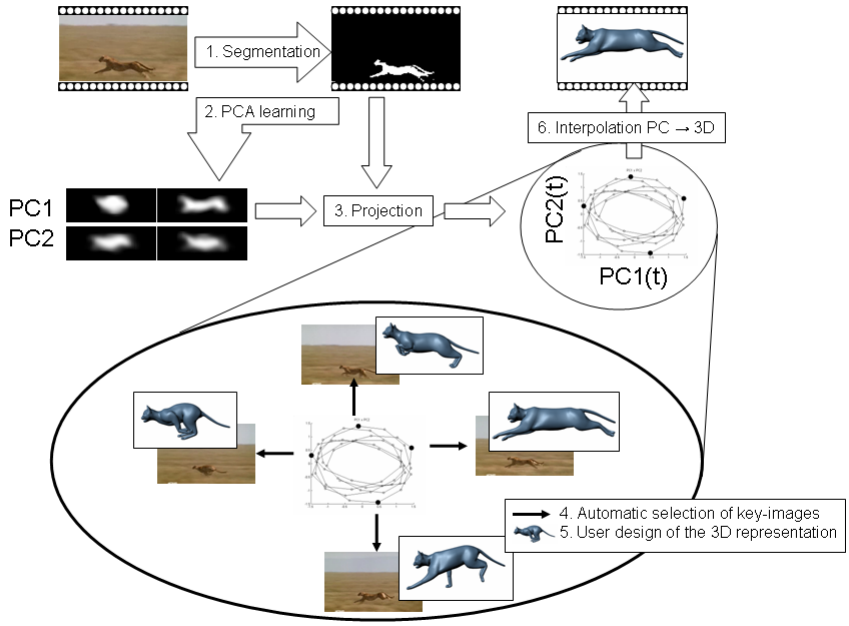

Figure 1: *Overview of the method*

1.1. Previous work

One of the first attempts to reconstruct animal motion from videos is Wilhelms's work [WG03]. Deformable contours (snakes) are used to extract the 2D motion of each limb's contour from a video sequence of a running horse. This motion is then transformed to 3D motion by matching 2D contours of limbs in the image with contours of limbs of a 3D model aligned on the image sequence. It is well known now that active contours methods are very sensitive to noise and have parameters which are difficult to tune. Wilhelms et al. [WG03] mention this problem and allow the user to reinitialize the active contours. This makes the method difficult to use in the general case, especially when limbs occlude each others. More generally, Gleicher et al. [GF02] show that current computer vision techniques for the automatic processing of videos fail to provide reliable 3D information such as stable joint angles over time. They conclude that using this kind of approach for the direct control of 3D animation at the joint level is currently not feasible.

Examples-based approaches have recently been recognized as a good alternative to traditional shape modeling and animation methods. The basic idea is to interpolate between a given set of examples, 3D pose or motion, mapped from an input parameter space to 3D data. Rose et al. [RBC98] parameterize the synthesis of new motion from motion capture data labeled with abstract parameters characterizing the style of motion. Lewis et al. [LCF00] interpolate shapes of a bending arm from joint angle values using Radial Basis Functions (RBF). They show that pose space mapping avoids well-known artifacts of traditional skinning methods. Sloan et al. [SIC01] extend this formulation by combining RBF and linear regression. All these approaches interpolate between well defined data - i.e. examples of 3D shapes or motion, labeled with user defined abstract parameters. Pyun et al. [PKC*03] show that a similar framework can be used to animate new faces from captured facial animation data. In this case, the abstract parameters are replaced by the 3D animation data that control the way the examples are interpolated over time. Visual features extracted from 2D images can also be used as input parameters to control pose space mapping. Bregler et al. [BLCD02] capture information from existing footage of 2D cartoon animation to control the blend shapes animation of 3D characters.

1.2. Overview

Our method is an example-based approach. We test video data as possible input parameters to control animation. Live video footage is challenging to process: because it lacks contrast and resolution, automatic feature extraction is not robust, and would require heavy user intervention. We rather convert the original images into normalized, binary images, on which Principal Component Analysis (PCA) is applied.

The images' coordinates in the Principal Component space provides an adequate set of parameters to control the 3D motion.

When input parameters are derived from a large set of data, all examples-based methods require that the user explicitly designate the examples. We propose a new and automatic criterion for selecting these examples. Radial Basis Functions (RBF) are used to interpolate between these pose examples over time, from the sequence of parameter values extracted from the video.

Section 2 presents our general pipeline for generating 3D animation from video: it details the chain of operations that we apply to the video sequences in order to extract adequate control parameters, and the way we interpolate given 3D pose examples to generate the animation. In particular, the conversion to binary images can either be fully automatic or use simple user input such as rough strokes sketched over the images: We show that both methods provide similarly good data for applying PCA.

Section 3 presents two extra contributions. First, we present a criterion for automatically selecting the best, minimal set of key-images from the video-data. Providing such a criterion prevents the user from spending hours carefully analyzing the input motion in order to find out which images he should associate with 3D pose examples. Second, we propose a simple algorithm to resolve ambiguities in the prediction of 3D pose from 2D silhouettes.

We validate our method in Section 4, by testing our approach on synthetic data: as our results show, we achieve a precise reconstruction of existing 3D animations of animal motion from video sequences of their 2D rendering, given that the right 3D shapes were associated with the automatically selected poses.

Section 5 presents our final results: wild animal motion is extracted from real life documentaries. Several features of our method, such as the option of filtering the coordinates in Principal Component space before applying the interpolation are discussed. We conclude and give directions for future work in Section 6.

2. Predicting 3D animation using PCA on images

2.1. Overview of the method

Our approach combines statistical analysis of binary images by PCA and animation by pose space mapping. The binary images can be generated by automatic segmentation. When automatic segmentation fails, we propose a sketching tool to label the video. In this case, white strokes on a black background create the binary image. PCA is then applied on the binary images, taking each image as a single observation vector. The projection coefficients of input images onto the Principal Components are analyzed to extract optimal examples of 3D poses to interpolate. These projection coefficients serve as input parameters to the pose space mapping interpolation and control the temporal evolution of the animation (Figure 1).

2.2. Reducing variability into binary images

Using PCA directly on images would encode any variation in appearance. In addition to variation due to motion, changes in illumination, camera motion and occlusion would be coded as well by PCA. Thus, before applying PCA, video images are segmented into binary images in order to filter such variation and isolate the foreground subject from the background. Assuming the user can provide some initial guess on the subject and background location on the first image by selecting two rectangular areas for each one on the first image of the sequence, a simple segmentation based on mixture of Gaussians can still provide accurate results (Figure 1 and top of Figure 2). This method is easy to implement and was sufficient for the purpose of our work on gaits generation. More elaborated techniques could be used and provide even more accurate input data to our approach [SM00].

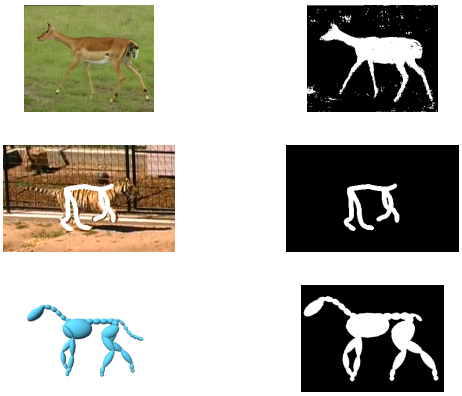

Figure 2: *Results of segmentation for our three sources of data: live video, sketching and synthetic*

When automatic segmentation fails, we propose to the user a sketching interface to label the video footage. The sketches does not need to be accurate as in Davis et al.[DAC*03], where the drawing needs to be precise enough so that the joints can be automatically recognized. In our case, the huge change in illumination and high occurrence of occlusions make impossible to claim for a careful joint to joint labeling. Instead, we rely on a raw labeling with strokes of the main features such as the spine, legs and head. It is not required to label every joint individually if they don't appear in the image. The idea is to similarly apply a PCA on images, either generated from segmentation or resulting from sketching.

Once the subject is isolated from the background, a region

of interest is automatically extracted around the silhouette by standard morphological analysis and detection of connected components. This process is applied to all the images in the video sequence. We keep track of the center of mass of the binary silhouette evaluated at the previous image so that the region of interest is still focussed on the correct connected component.

This step allows us to get rid of variance due to camera motion which is not relevant to the true motion of the tracked subject. Unfortunately, it also filters out the vertical translation of the animal, which is relevant to motion. Nevertheless, we are not trying to extract the translation as an independent parameter. Instead, our aim is to capture the overall timing and predict animation by interpolation of 3D pose examples. Consequently, if a vertical translation is set in the pose examples, assuming such a motion is correlated with the rest of the visible motion in the images sequence, it will appear in the final animation. Typically, a full body extension is correlated with a flight part in the gait scenario.

From this pre-process, we end up with a sequence of binary images. We give here the data specifications for our 7 test sequences in terms of number of frames, size in pixels of the original image, and size in pixel of the tracked window of the silhouette.

Sequence	Number of frames	Original image	Binary image
Horse walk	100	320x240	320x240
Horse canter	100	320x240	320x240
Horse gallop	100	320x240	320x240
Cheetah run	137	192x144	90x34
Tiger run	60	720x480	448x216
Antilope walk	122	352x288	232x173
Giraffe walk	73	352x288	195x253

2.3. Principal Components as input visual features

Principal Components Analysis (PCA) is a standard method for data reduction. It consists in finding a set of principal axes, linearly spanning the variance of multidimensional data. Turk and Pentland introduced one of the first implementations of PCA on images to perform face recognition (eigen-faces) [TP91]. In this case, each image is considered as an independent observation where all the pixels values are stacked in a single vector. Eigen-images have been widely used to reduce, classify and recognize regular patterns from images. As a new contribution we show that PCA on images can encode variation due to motion only and can be used not only to classify shapes but also to continuously predict change in motion. We will take benefit of this property in the interpolation scheme.

PCA consists in calculating the eigenvectors and eigenvalues of the covariance matrix of the collected data. In our case, each rectangular image of the sequence is viewed as a row vector $\mathbf{i}(t)$ of all the pixels values stacked together. We gather all the n images over a sequence in a matrix $\mathbf{I}$, after having subtracted the mean image $\bar{\mathbf{i}}$:

$$\bar{\mathbf{i}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{i}(t) \tag{1}$$

$$\mathbf{I} = \left[\left(\mathbf{i}(t_1)-\bar{\mathbf{i}}\right)^t,\ldots,\left(\mathbf{i}(t_n)-\bar{\mathbf{i}}\right)^t\right]^t \tag{2}$$

The PCA is then formulated as:

$$\frac{1}{n}\mathbf{I}^t\mathbf{I}\mathbf{E} = \mathbf{E}\mathbf{D} \tag{3}$$

$$\mathbf{E}^t\mathbf{E} = \mathbf{1} \tag{4}$$

Finally, we take as input vector of the animation the projection coefficients onto the Principal Components stacked as column vectors in matrix $\mathbf{E}$ and normalized by the square roots of the eigenvalues stacked in the diagonal matrix $\mathbf{D}$:

$$\mathbf{p}(t) = \left(\mathbf{i}(t)-\bar{\mathbf{i}}\right)\mathbf{E}\sqrt{\mathbf{D}^{-1}} \tag{5}$$

We recapitulate below the results of PCA in terms of part of the variance covered by each Principal Component with respect to the total variance of the data for our 7 test sequences.

Sequence	PC1	PC2	PC3	PC4
Horse walk	33.7	23.7	11.4	8.56
Horse canter	32.5	14.5	9.17	8.78
Horse gallop	31.1	19.9	11.0	8.33
Cheetah run	44.7	11.6	9.93	7.79
Tiger run	15.2	10.5	6.14	4.69
Antilope walk	21.5	12.2	8.40	6.91
Giraffe walk	42.8	15.8	11.1	5.63

2.4. Interpolation

Our goal is to generate animation parameters (position and joint angles) $\mathbf{x}(t)$ from the values of projection coefficients $\mathbf{p}(t)$ computed from PCA. We use interpolation of m 3D pose examples $[\mathbf{x}(t_i)]_{i=1\ldots m}$, corresponding to m images in the video sequence for which we know the projection coefficients $[\mathbf{p}(t_i)]_{i=1\ldots m}$ at time t_i in the video sequence. For clarity, we note $\mathbf{x_i}$ and $\mathbf{p_i}$ for respectively $\mathbf{x}(t_i)$ and $\mathbf{p}(t_i)$.

Three main methods for scattered data interpolation are used in example-based method approaches: linear interpolation[BLCD02], Radial Basis Function [LCF00] or a combination of both[SIC01]. In the latter case, linear interpolation allows us to cope with cases where input data could be sparse and require a stable behavior for extrapolation. In our case, input data is the results of PCA and as such is already linearly compact. For this reason, Radial Basis Function (RBF) were enough to deal with our case. This

general interpolation scheme is formulated as linear combination of distance functions $h(r)$ (the RBF) from m interpolation points in the input space:

$$\mathbf{x}(\mathbf{p}) = \sum_{k=1}^{m} h(\|\mathbf{p} - \mathbf{p_k}\|)\mathbf{a_k} \tag{6}$$

where $\mathbf{p}$ is the input vector and $\mathbf{x}$ the predicted vector. $h(r)$ are the RBF. $\mathbf{a_k}$ are unknown vectors to be determined. If the RBF are stacked into a single vector $\mathbf{h}(\mathbf{p})$ and the unknown coefficients $\mathbf{a_k}$ as row vectors into a matrix $\mathbf{A}$, we have the formulation :

$$\mathbf{x}(\mathbf{p}) = \mathbf{h}(\mathbf{p})\mathbf{A} \tag{7}$$

$$\mathbf{h}(\mathbf{p}) = [h(\|\mathbf{p} - \mathbf{p_1}\|) \dots h(\|\mathbf{p} - \mathbf{p_m}\|)] \tag{8}$$

$$\mathbf{A} = \left[\mathbf{a_1^t}, \dots, \mathbf{a_m^t}\right]^t \tag{9}$$

As interpolation points, we use m 3D pose examples $\mathbf{x_i}$ and the the values of the m associated input parameters $\mathbf{p_i}$ of the corresponding key-image. $\mathbf{A}$ has to be solved so that $\|\mathbf{x}(\mathbf{p_i}) - \mathbf{x_i}\|$ is minimal. This minimization in a least square sense leads to the standard pseudo-inverse solution :

$$\mathbf{A} = \left(\mathbf{H^t H}\right)^{-1} \mathbf{H^t X} \tag{10}$$

where,

$$\mathbf{X} = \left[\mathbf{x_1^t}, \dots, \mathbf{x_n^t}\right]^t \tag{11}$$

$$\mathbf{H} = \left[\mathbf{h}(\mathbf{p_1})^t, \dots, \mathbf{h}(\mathbf{p_m})^t\right]^t \tag{12}$$

The final formulation is then :

$$\mathbf{x}(\mathbf{p}) = \mathbf{h}(\mathbf{p}) \left(\mathbf{H^t H}\right)^{-1} \mathbf{H^t X} \tag{13}$$

Note that this can be re-formulated exactly as an interpolation of the $\mathbf{x_i}$:

$$\mathbf{x}(\mathbf{p}) = \sum_{i=1}^{m} w_i(\mathbf{p})\mathbf{x_i} \tag{14}$$

by extracting the matrix $\mathbf{h}(\mathbf{p}) \left(\mathbf{H^t H}\right)^{-1} \mathbf{H^t}$. In [AM00], Alexa et al. compress and animate 3D sequences from principal components (PC) learnt on a fully available sequence of 3D data. In our case, PC are learnt from image space and animation of 3D data is controlled by interpolation.

The value of $\sum_{i=1}^{m} w_i(\mathbf{p})$ should stay close to 1 to guarantee that a point $\mathbf{p}$ in the input space is close enough to interpolation points and any $w_i(\mathbf{p})$ should be close to $[0,1]$ so that $\mathbf{x}(\mathbf{p})$ stays close to the convex hull of the 3D pose examples.

For the choice of $h(r)$, a common practice is to use a gaussian function for its $\mathcal{C}^\infty$ continuity properties:

$$h(r) = e^{-\alpha r^2} \tag{15}$$

The parameter α in equation 15 needs to be determined. Statistically, projections on PC are homogenous with standard deviation. This means data will be spread approximatively in every projected direction over the same interval $[-1;+1]$ - varying according to the nature of distribution. Assuming interpolation points are well spread, we take a value of 2 as a raw estimate of the distance between interpolation points. At midpoint between two interpolation points, we expect an equal influence. This can be translated into the fact that we want $h(r)$ to be equal to 0.5 when $r = 1$. This leads to an estimate of $\alpha = \ln 2$.

All previous works on example-based animation rely on the user to decide where 3D pose examples need to be provided [LCF00, SIC01, PKC*03]. In our case, this would mean selecting key-images among thousands of a video sequence. Given the number of key-images to provide, we present an automatic criterion to select these ones within the video sequence.

3. Key-images selection

3.1. Criterion for automatic selection

We want smooth mapping between the image space and the animation space as we based all our timing control on images. A small change in the image space must produce a small change in the animation space. We notice that the interpolation scheme on RBF involves the inversion of a matrix $\mathbf{H^t H}$, build from the interpolation points, as it has been shown in the previous section. Consequently, to ensure a stable interpolation, and thus a smooth animation, *we select key-images over the sequence which minimize the condition number of the matrix* $\mathbf{H^t H}$ *to invert*. The condition number is evaluated as the ratio of the largest singular value of the matrix to the smallest. Large condition numbers indicate a nearly singular matrix.

This criterion is generally applicable to any example-based method. It can be used to select any number of input examples, key-images in our case, when they have to be chosen within a large set of data. The singular values of $\mathbf{H^t H}$ are the squared singular values of $\mathbf{H}$. This matrix measures the respective distances between the interpolation points. Intuitively, the criterion on condition number thus selects input examples which are equally spread within the data set. Having all the singular values closed to each other means they equally sample every direction of the input space.

In practice, as will be shown in section 4 and 5, only few principal components and few 3D pose examples are needed. This allowes us to implement a simple combinatory approach for the condition number criterion: for each sequence of n frames, given a number of c principal components to consider and a number m of key-images to select, we evaluate the condition number of all the $\binom{n}{m}$ matrices $\mathbf{H^t H}$. The $\mathbf{H^t H}$ matrix is square and its dimension is m. We keep

the set of m key-images within the whole sequence providing the $\mathbf{H^t H}$ matrix having the smallest condition number. Keeping only a few Principal Components makes the computation fast. We tested with up to 5 Principal Components, but experiments showed that 2 were enough as will be detailed in following sections.

As an example, for the prediction of 3 sequences of animation from synthetic images, we plot the projections on the two first components as a 2D graph and search for the best 4 examples based on the condition number criterion (next section will show that 2 PC and 4 keys is the best configuration for the prediction of this specific gait). In this case the condition number criterion has the particularity to select examples at approximatively the extreme variation of the two first PCA projections (Figure 3).

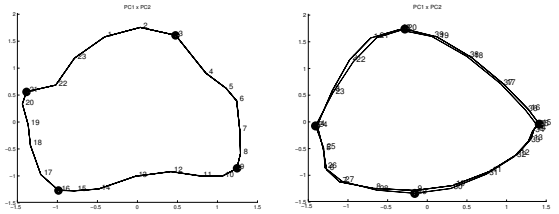

Figure 3: *PC1(t)xPC2(t) : PC projections across time for two synthetic sequences of horse: canter and walk. Frames are numbered for one cycle. Circles are selected examples by the condition number criterion.*

Intuitively, the more key-images are given, the better the interpolation will be. As any key-image will require the user to provide a 3D pose example, a compromise must be found. The question of the number of key-images needs to be examined on a case-to-case basis. From our experiments on animal gaits, we observed good results with 4 pose examples for the running cases and 8 pose examples for the walking cases.

3.2. Resolving 2D ambiguities with switching models

At this point, our method predicts 3D motion from silhouette images. It results in a unique 3D pose for each distinct input image. In some cases however, two different 3D poses can lead to very similar silhouettes when viewed from the side (Figure 4). This is very common in motions that consist in a succession of two symmetric phases, such as quadrupeds walking. The motion predicted by RBF still provides good results but only on one half of the period of the original gait.

To avoid this problem, it is first necessary to provide two different 3D pose examples for each of the ambiguous silhouette of the key-image and secondly to build a method to correctly choose between these two poses during the generation of the 3D animation.

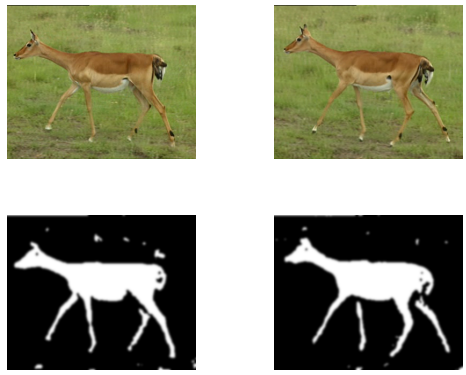

Figure 4: *Two different poses can produce similar silhouettes.*

We solve for the first problem with a simple algorithm:

1. We select m initial key-images with the standard method and build the animation by associating 3D poses to key-image and using RBF prediction. If the user acknowledges issues about pose ambiguities, we go to step 2.
2. For each key-image, we automatically search for its closest image in the PCA space and propose it to the user as the alternative pose for this silhouette. We constrain this image to be at least 3 frames further than the initial key-image to guarantee that we are in another half-cycle.
3. When the user validates the proposed image as the key-image corresponding to the same silhouette but at a different pose, we ask the designer to provide the appropriate 3D pose example.
4. We iterate until each of the m initial key-images of step 1 has its associated key-image corresponding to the opposite pose.

At the end of this process, we have doubled the number of m initial key-images and corresponding 3D pose examples. Figure 5 provides an example for this algorithm with $m = 4$ initial key-images. We are able now to generate a full cycle of motion. To generate animation, the same method of prediction from images is kept, but instead of keeping the same m 3D pose examples, we switch between q sets of m 3D pose examples as time evolves, taken from the $2m$ 3D pose examples selected by the previous algorithm. We call these q sets the switching models. The prediction of animation parameters is extended as follows :

$$\mathbf{x}(\mathbf{p}_t) = \sum_{k=1}^{q} w_{\sigma_k(s_t)}(\mathbf{p}_t)\,\mathbf{x}_{\sigma_k(s_t)} \tag{16}$$

$$s_t = switch(\mathbf{p}_t, s_{t-1}) \tag{17}$$

where s_t represents a phase state index in term of switching model, $\mathbf{x_i}$ the $2m$ pose examples, and w_i the model weight given the input image and the current phase state. The function $switch(\mathbf{p}, s)$ indicates which set of m pose examples

needs to be used in the prediction algorithm. It is a discrete state variable, incremented each time we detect that we have reached the last 2D silhouette within a set of m key-images. The change of silhouette in key-images is easily detected by a distance function in the PCA space.

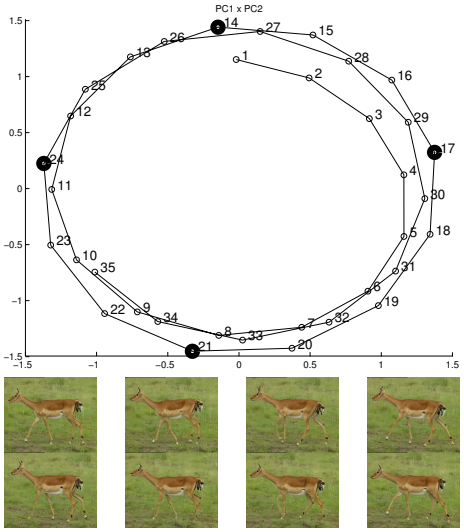

Figure 5: *Selecting more key-images to resolve ambiguities in the PCA space. Images 14, 17, 21 and 24 are selected as initial key-images. In a second step, 27, 4, 8 and 11 are selected as candidates for ambiguous pose, based on their coordinates in the PCA space. First row : frames 14, 17, 21 24; Second row : frames 27, 4, 8, 11.*

Switching between $q = 2$ models of m pose examples would allow to explore the whole animation space, as we have just doubled the number of initial m pose examples. However, the transition between two models turned out to be unstable. We solved this problem by introducing overlaps between intermediate models. The use of $q = 4$ switching models allows smooth transitions. In practice we use $m = 4$ pose examples to describe half of a cycle motion. The function $\sigma_k(s)$ gives the 4 indices of pose examples at 4 state positions with 2 overlapping pose examples between two consecutive steps:

$$\sigma_i(j) = \sigma_{ij} \tag{18}$$

$$(\sigma_{ij}) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 5 & 6 \\ 5 & 6 & 7 & 8 \\ 7 & 8 & 1 & 2 \end{pmatrix} \tag{19}$$

4. Validation on synthetic images

We have validated our method by taking as input images the rendering of a skeleton horse 3D model. Joints are represented as ellipsoids (Figure 2). The choice of such a model was made to get rid of any bias that a skinning algorithm would introduce. We report results for three sequences: gallop, canter and walk. By using synthetic images, we can still test the full pipeline as described in section 2. In addition, we can compare with the original animation parameters. This evaluation gave use hints on the number of principal components and examples that should be used.

We have exhaustively evaluated the results using an increasing number of key-images (starting at two) and an increasing number of PC (starting at one). Given the number of key-images to select from the video, the condition number criterion tells what key-images to select. The corresponding 3D pose examples are provided by the original animation sequence. We evaluate the results by computing the mean (and standard deviation) of the absolute difference over all the joint angles for the main rotation axes (perpendicular to the image plane, 36 angles in the case of our model) between original and predicted values.

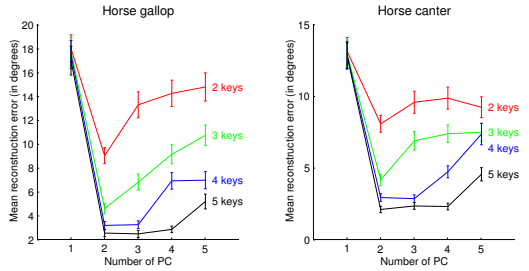

Figure 6: *Evaluation for the gallop and canter sequences. Each curve corresponds to the mean error with a fixed number of examples (2 to 5), with respect to the number of components used as input parameters.*

From Figure 6, we immediately observe that adding the third and following components introduce noise. This suggests they are coding information not relevant to the gait motion. As for the number of keys, as expected, the more examples are provided, the smaller the error. A good compromise arises on 4. Adding a fifth keys decreases the mean error less than a degree. The results are confirmed on the video provided with this paper. Two or three pose examples, although optimally selected by the condition number criterion, are not enough. With four 3D pose examples and two Principal Components, we obtain a very good match between the original animation and the predicted animation from images.

5. Processing live video sequence

We discuss now how to apply our approach on live video images, sometimes emphasized by a rough sketch as mentioned in section 2. As detailed below, strictly focussing on the first two PC and applying a band-pass filter to the PC trajectories along time enables us to achieve as good visual results as with the synthetic data.

5.1. Restricting to the two first PC

In the case of the synthetic examples, the first two components exhibit consistently interpretable behavior. For example, for the gallop of the horse, The first component (PC1) encodes a variation between a flight phase, when none of the feet touch the ground, and a grouped phase. The second component (PC2) corresponds to an opposition between a rising phase, when the horse jumps off the ground, and a descending phase, when the horse front feet hit the ground (Figure 7).

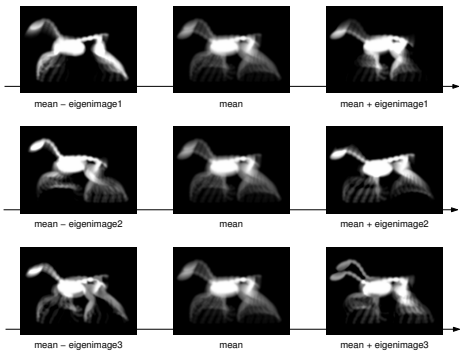

Figure 7: *Variation encoded by the first three eigenvectors for the horse gallop sequence. Middle column is the mean shape, each row corresponds to the variation along an eigenvector.*

Numerical evaluation on synthetic images have suggested that the two first PC are optimal to achieve good prediction. Image segmentation and sketching by hand will naturally introduce more noise in PC curves, making PC unstable and poorly reliable to predict relevant motion. We decide from these observations that the only 2 first PC should be kept for live video and sketched images.

We confirm this hypothesis on the cheetah sequence where similar interpretation as that for the horse gallop can be made on the two first principal components (Figure 8).

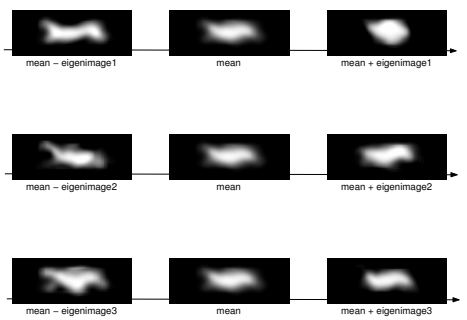

Figure 8: *Variation encoded by the first three eigenvectors for the cheetah sequence.*

5.2. Spectrum regularization

On the synthetic sequence, the time variation of the projections on the first two components shows a shift in phase of one fourth of the cycle period, corresponding to an alternation of jump, flight, landing and grouping legs. This produces the circular pattern shown on Figure 3. This configuration has been reproduced on every examples of synthetic images. Consequently, we adopt the configuration of projections on PC1 and PC2 in a circular pattern as a characterization of a video sequence to be usable with our method. In the Fourier domain, this configuration corresponds to peaks at the same location for projections on PC1 and PC2, and a phase difference of approximatively $\frac{\pi}{2}$.

Live video can thus be diagnosed as not usable by our method if it does not have projections on PC1 and PC2 staying within a certain bandwith that we automatically estimate. The first component encodes most of the variance and is considered to be representative of the fundamental cyclic variation. Its spectrum will thus be centered around a frequency corresponding to the period of the cycle. All our experiments confirm this hypothesis. From a Fourier Transform we get the frequency of maximum amplitude. We fit a peak function centered on this frequency of the form:

$$peak(f) = \frac{1}{1 + (\frac{f - f_0}{f_b - f_0})^2} \tag{20}$$

f_0 is set at the frequency of maximum amplitude and f_b is set so that it corresponds to the closest frequency to f_0 having an amplitude of half of the maximum. We deduct a bandwidth of $[-3(f_b - f_0); +3(f_b - f_0)]$, corresponding to end points at 10% of the maximum amplitude (Figure 9).

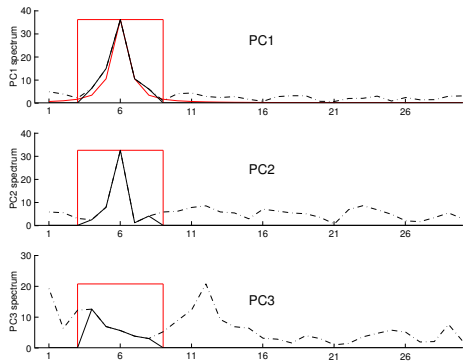

Figure 9: *Spectrum of the 3 first PC of the cheetah spectrum. The peak function is fitted to PC1, and a rectangular window is deduced.*

The second component is filtered by the same band-pass filter. What is expected is that the second component shows a similar peak, creating the circular pattern. We can evaluate how this hypothesis is respected by comparing how

much the reconstructed signal after filtering, matches the original signal. For this, we compute the correlation coefficient between the original principal component signal and the filtered principal component signal. We have observed that for our test sequences, the two first components shared the same peak ($r \geq .6$), while the following components do not ($r \leq .3$), see table below. This provide a numerical criterion to evaluate if our method can be successfully applied to a video sequence as we have presented it.

PC:	PC1	PC2	PC3	PC4	PC5
horse gallop	.90	.92	.01	<.01	<.01
horse canter	.94	.89	.08	<.01	.01
horse walk	.99	.94	<.01	.05	<.01
cheetah run	.91	.61	.14	.21	.25
antilope walk	.78	.91	.14	.10	.06
tiger run	.87	.72	.26	.18	.23
giraffe walk	.92	.82	.24	.19	.28

Figure 9 shows the results in the Fourier domain for the cheetah sequence which conforms to the criterion. When the live video sequence fails to conform to this criterion, we suggest using the sketch approach. If the sketch approach still fails to meet the criterion, we diagnose that our method cannot work on the analyzed video.

5.3. Results

We show in the video provided with the paper results for a cheetah run, an antilope walk (both automatically segmented) and a tiger run and a giraffe walk where automatic segmentation fails but sketch images succeed.

Figure 10 shows the evolution of the weights of the four 3D pose examples for the cheetah sequence. We have an exact interpolation at these pose examples. For the rest of the sequence, we observe a correct generalization, the influence of each pose example appears at a right pace in a coherent order. Note that the sum of weights stays close to one, guaranteeing that the input parameters are always close to an interpolation point. The weights are sometimes outside of the range of $[0,1]$ as they are not constrained in the RBF formulation. This lets the resulting pose leave the convex hull of the pose examples. This flexibility allows some extrapolation in 3D space introduced by image variations along the sequence. A control could be easily added to maintain these weights within a safe range in order to avoid the generation of strange pose, too far outside of the convex hull of the pose examples.

Finally, Figure 11 gathers the final results about key-images selection. It shows the automatically selected key-images and the associated 3D pose provided by the artist. The full video of all the tested sequences are given in the demo movie file.

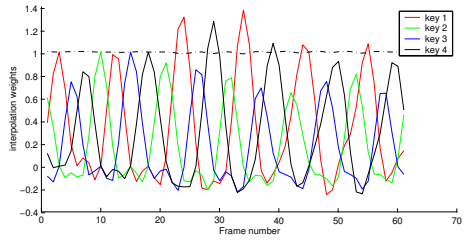

Figure 10: *Evolution of the weights of the 4 examples for the cheetah sequence. The dashed line is the sum of weights.*

6. Conclusion

When traditional motion capture of non-cooperative subjects such as wild animals is not feasible, live video footage can still provide significant information on motion. This paper has proposed novel and robust tools to analyze video data and make it applicable to the generation of 3D motion. We rely on Principal Component Analysis to extract parameters from binary version of the input images. As our result show, a small number of parameters is sufficient for cyclic animal gaits: using the two principal components already gives good results. We provide a criterion for selecting the best set of key-images from the video. In our application, the selected poses can easily be interpreted in terms of extremal images in the 2D Principal Component space.

Our work shows that Principal Component Analysis (PCA) can be applied onto a sequence of 2D images to control 3D motion. PCA on images helps to give a quantification of the significant changes in the appearance of the video. The RBF interpolation of pose examples aims at transposing the pace of video changes into the animation domain. The automatic selection of examples helps to focus the effort of the designer on the most important key-frames.

As a future work, we are planing to explore non uniformly cyclic motion such as transition between gaits and the addition of physically-based constraints to animate non-cyclic part of the motion. We are also studying how to re-use existing PCA basis and its 3D associated poses to automatically analyze a new video sequence thanks to morphological adaptation in the image space.

References

[AM00] ALEXA M., MÜLLER W.: Representing animation by principal components. In *Proc. EUROGRAPHICS'00* (2000).

[BLCD02] BREGLER C., LOEB L., CHUANG E., DESHPANDE H.: Turning to the masters: motion capturing cartoons. In *Proc. SIGGRAPH'02* (2002).

[DAC*03] DAVIS J., AGRAWALA M., CHUANG E.,

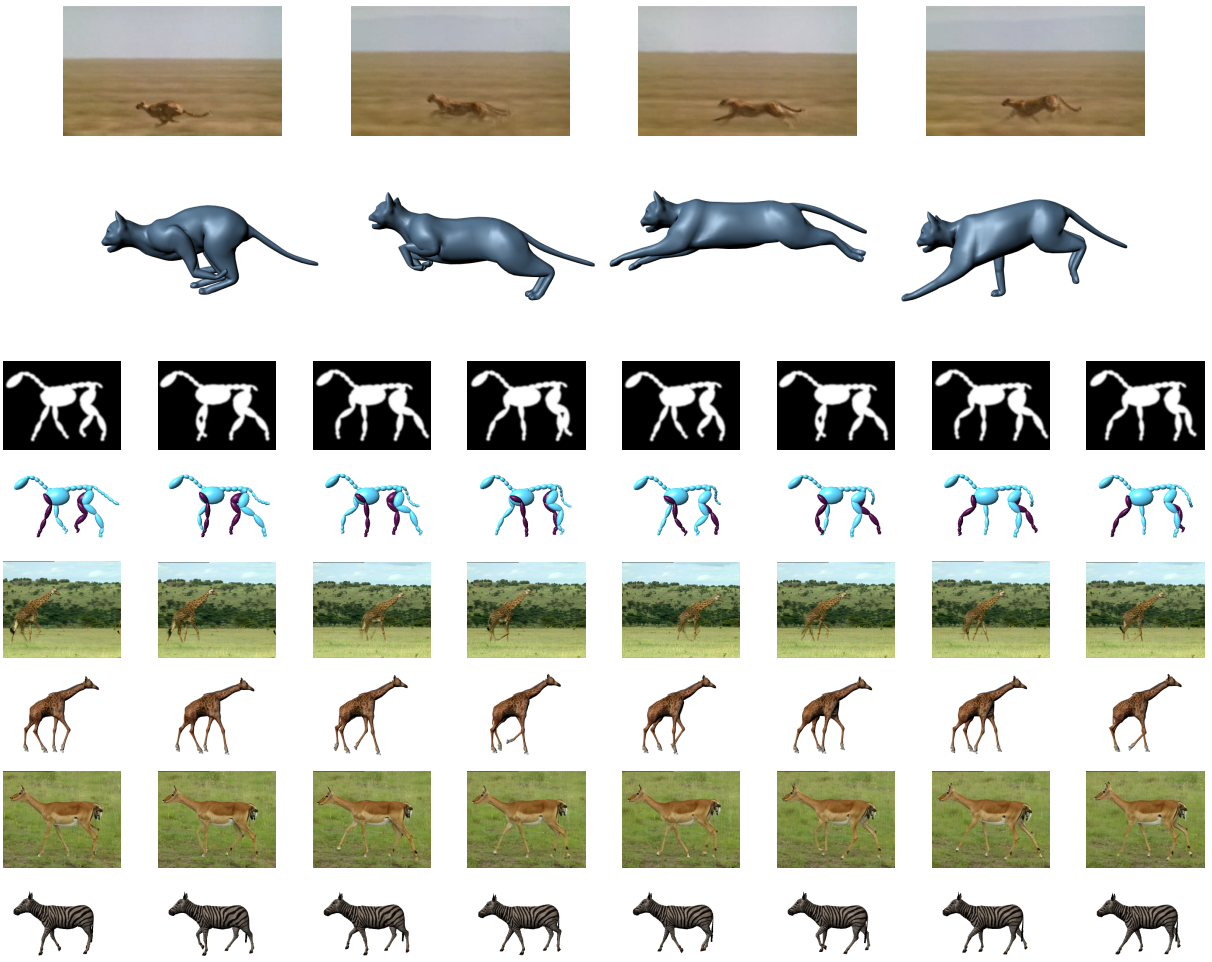

Figure 11: *Selection of key images from video sequences.*

POPOVIC Z., SALESIN D.: A sketching interface for articulated figure animation. In *Proc. EG/SIGGRAPH Symposium on Computer Animation, SCA'03* (2003).

[GF02] GLEICHER M., FERRIER N.: Evaluating video-based motion capture. In *Proc. of Computer Animation, CA'02* (June 2002).

[LCF00] LEWIS J., CORDNER M., FONG N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven defomation. In *Proc. SIGGRAPH'00* (2000), pp. 165–172.

[PKC*03] PYUN H., KIM Y., CHAE W., KANG H. W., SHIN S. Y.: An example-based approach for facial expression cloning. In *Proc. EG/SIGGRAPH Symposium on Computer Animation, SCA'03* (2003), pp. 167–176.

[RBC98] ROSE C., BODENHEIMER B., COHEN M. F.: Verbs and adverbs: Multidimensional motion interpolation using radial basis functions. *IEEE Computer Graphics and Applications 18*, 5 (Sept. 1998), 32–40.

[SIC01] SLOAN P.-P. J., III C. F. R., COHEN M. F.: Shape by example. In *Proc. I3D'01* (2001).

[SM00] SHI J., MALIK J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8 (2000).

[TP91] TURK M., PENTLAND A.: Eigen faces for recognition. *Journal of Cognitive Neuroscience 3*, 1 (1991).

[WG03] WILHELMS J., GELDER A. V.: Combining vision and computer graphics for video motion capture. *The Visual Computer 19*, 6 (Oct 2003).