EUROGRAPHICS '10 / N.N. and N.N.
(Editors)

Volume 0 (2010), Number 0

The Virtual Director: a Correlation-Based Online Viewing of Human Motion

paper 1169

Abstract

Automatic camera control for scenes depicting human motion is an imperative topic in motion capture base animation, computer games, and other animation based fields. This challenging control problem is complex and combines both geometric constraints, visibility requirements, and aesthetic elements. Therefore, existing optimization-based approaches for human action overview are often too demanding for online computation.

In this paper, we introduce an effective automatic camera control which is extremely efficient and allows online performance. Rather than optimizing a complex quality measurement, at each time it selects one active camera from a multitude of cameras that render the dynamic scene. The selection is based on the correlation between each view stream and the human motion in the scene. Two factors allow for rapid selection among tens of candidate views in real-time, even for complex multi-character scenes: the efficient rendering of the multitude of view streams, and optimized calculations of the correlations using modified CCA. In addition to the method's simplicity and speed, it exhibits good agreement with both cinematic idioms and previous human motion camera control work. Our evaluations show that the method is able to cope with the challenges put forth by severe occlusions, multiple characters and complex scenes.

1. Introduction

Human motion animations in games and 3D environments applications are ubiquitous and continue its rapid growth. As a result, the necessity for motion-sensitive camera handling techniques is growing respectfully. Controlling the camera to generate video clips which expresses well the motion is challenging since the problem poses many requirements with conflicting constraints. Several studies had proposed systems specializing on human motion data, which demonstrated encouraging results; however such solutions are offline methods or can only be applied for simple scenes.

Camera control methods that support human motion apply various optimization techniques, searching for an optimal camera control path subject to many constraints including general ones which hold for any scene such as collision avoidance, occlusion detection, coherency and camera smooth and slow-moving trajectory, and constraints specific for human action such as significance of gestures, internal occlusions and different significance of limbs throughout the motion [DZ94, BTM00, HHS01, MK06, CO06, ACO*08].

An alternative to the optimization methods are rule based methods which rely on cinematopgrahy guidelines referred to as "cinematic idioms" formulated in the early days of the cinamtography theory. As basis for the theory, the founders of cinamtography had suggested using multiple cameras which are orchestrated together using a known set of cinematic idioms rules. These guidelines were designed to introduce simple scene configurations describing the scene events, in a coherent manner. For example, in a conversation scene, the discussing parties are to be viewed by the cameras to provide the viewer the feeling of taking part in the conversation [Ari76].

The cinematic idioms, which describe the preset camera configuration and their switching strategy, were suggested for a certain set of specific scenarios, and were implemented by several of the early camera control systems [HCS96], however, the extension of these idioms to other scenes and scenarios were not pursued. The cinamatic rules abstraction suggested by several camera control methods use geometric constraints which are easy to compute [DZ94, HO00, CO06]. However they are limited and insufficient for expressing the details of human motion based scenes as much as optimization methods [ACCO05, ACO*08].

In this work, we introduce an effective automatic method which is both fast enough to allow online computation and generic in a sense that it can be applied to arbitrary human

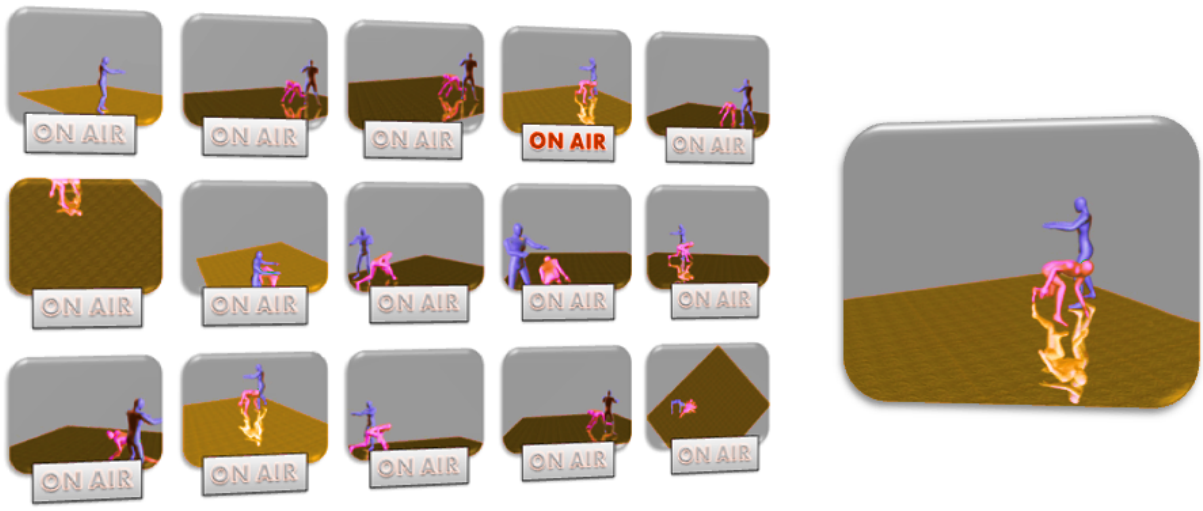


Figure 1: *Our method generates a coherent presentation of human motion animation scenes by automatically evaluating the view streams of a multitude of virtual cameras and constructing the sequence of the most suitable streams in realtime.*

action scenes. The key idea is to use a multitude of virtual cameras from which at any given time the best view is selected. The premise is that a rather modest number cameras is dense enough so that the view from at least one of them is expressive enough. Our method is based on a fast measurement of the correlation between the given scene animation and the camera output. This measurement allows to assess how descriptive each view is with respect to the source animation, and select the best view which captures the essence of the original animation. We further extend this notion of correlation to define a strategy for determining instants in time where a switch between the available virtual cameras can take place, thus introducing multi-shot scheme. Our technique in a sense mimics the professional solution of real-time camera crew, e.g., the teams which cover online TV sport events. In such teams, the camera control director primarily selects the best camera view stream at each given time.

As we highlight in our novel approach, picking the best view is by far computationally easier than optimizing the camera pass over the whole space. Nevertheless, as we demonstrate, the online video generated by the multitude of camera is expressive enough and approximates well the quality of an offline optimization method. Furthermore, the simplicity of our method allows handling of complex scenes under severe occlusion conditions as well as multi-character scenes, thus extending the applicability to new cases which were not handled by previous approaches.

The presented method introduces several additional contributions which we examine in this work: first, we show that the view-animation correlation can be effectively used to define a partial order between the various view streams. This partial order, under some restrictions, can be used to compose video summaries of scenes even without the animation data. Secondly, we show that the scheme fits existing camera idioms, and therefore can be considered as an extension and generalization of the "spirit" of the defined idioms to different scenes and scenarios. Lastly, the simplicity of the method can be used in various applications, such as autonomous game viewing cameras, quick overview generation for 3D authoring tools, etc.

The following section describes some of the recent advances in camera control and human motion sensitive camera control, next we describe the details of our method in Sections 3, 4. We discuss our results, show a comparison to both previous work, and present a user study which assess our results in Section 5. We conclude with a summary in Section 6.

2. Related work

The problem of camera control focuses on searching for a suitable camera configuration for capturing a scene narrative, while obeying a set of cinematographic rules [Mas65]. The problem is challenging since the camera configuration is subject to numerous constraints, such as occlusion, objects visibility, and layout in the resulting image [GW92, CMN*05]. A survey of the different approaches to the problem is described in Christie and Olivier overview [CO06].

The art of cinematography placed a special attention on action scenes with single and multiple actors. Guidelines were established, defining a set of standard principles of camera configurations and transitions [Ari76, Mas65, Kat91] often referred to as *cinematography idioms* such as estab-

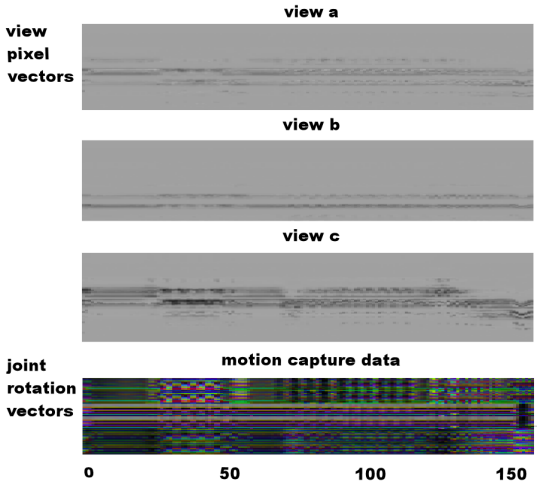


Figure 2: *An example of three input views (top three strips), and animation stream (bottom strip). Each frame is presented as a single column in strips. The view columns include the view pixels vectors and the animation data includes the pose joints details. Our method automatically seeks for correlation between the view and the animation data, and uses the correlation to select the best view.*

lishing the scene configuration, avoiding jump cuts, avoiding crossing the scene action line and other rules. An early attempt to mimic some of the idioms in a complete framework is the "The Virtual Photographer" [HCS96]. In their work they use state machine control together with predefined camera configuration to implement an idioms based camera control system. In general, methods for camera control had handled various soft and hard idioms related constraints, based on geometric properties of the camera position, and on the attributes of captured view (e.g., [DZ94, HCS96, BTM00, HHS01, LST04, MK06]).

Most of the existing camera control systems are generic and are not designed to specifically support or express human actions. Only recently, several studies are addressing this problem with specialized methods for human motion. Kwon and Lee [KL08] introduce a camera control technique for character based scenes. Their approach is based on the measuring the *motion area*, that is, the integrated area spanned by the character bones motion. Their work focus on a selection of static camera positions which are then extended into a camera path by interpolation. In cases of rapid actions, it is shown that the method generates a non-smooth and unpleasing camera results. Assa *et al.* [ACO*08] describe an optimization method with several constraints specifically tuned to the description of human action, such as self occlusion and action saliency. They use a global optimization to generate a suitable camera path, with possible splitting the scene-capturing to multiple shots. Their work is based on quantum annealing and cannot provide an online solution.

The intelligent galleries [VBP*09] uses supervised learning to realize good camera positions for various scenes. Their approach, although not specifically designed to human action, suggests that a single general framework such as machine learning analyzing the camera views can be used to define camera preferences in many different domains. Similarly, in this work we show that the general tool of motion correlation can by applied in an unsupervised manner to human aware camera control solution.

Although the effectiveness of using multiple shots in scenes is known since the dawn of cinematography, not many camera control systems had utilized this tool. A brief review of existing camera control studies shows that only a small number of the work handles multiple shots. These studies typically employ shots to solve cases of camera-object collisions, or cases in which the camera view quality deteriorates as a result of a rapid scene action coupled with slow camera movement [DZ94, HO00, CO06, ACO*08].

3. Views correlation

Our goal is to present a dynamic scene by using a sequence of views that capture as much of the scene action as possible. Given a set of view streams, we would therefore like to prioritize them according to how well they express a given animation stream. We use a simple view representation, which employs for each frame the vector of all pixel values. As we show in this work, even low resolution rendering of the views convey sufficient information for the purpose of our processing. The animation stream consists of motion capture data joint details, such as their relative angles. Note that even the low resolution frames and straightforward animation representations used, produce large input data streams. A single frame of a typical view stream is of 9,000D. This large amount of data, illustrated in Figure 2, require efficient computational means.

While the view and the animation representations differ in their properties, both reflect information arising from the underlying motion, and are therefore correlated. The premise is that expressive views have higher correlation with the animation stream. Moreover, if the view contains a significant amount of other types of variability that are caused by camera motion or by the motion of background objects, the correlation diminishes.

3.1. Efficient computation of CCA

The correlation between the view stream vector and the animation stream vector is computed via the Canonical Correlation Analysis (CCA) technique. CCA is a statistically solid tool that is frequently used in various data analysis [Bor98]. It was introduced by Hotelling as a method of examining

the dependencies between two sets of random variables. The technique finds two linear transformations such that the correlation between the transformed pairs of variables is maximal.

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be sets of real-valued vectors with zero mean (if they are not, the mean is subtracted first) such that $x_i \in \mathbf{R}^{d_X}$ and $y_i \in \mathbf{R}^{d_Y}$. Let $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ be matrices with x_i and y_i in their i'th column, respectively. Given a dimension d, CCA finds matrices W_X of dimensions $d_X \times d$ and W_Y of dimensions $d_Y \times d$ that maximize the element-wise correlations of the transformed vectors $W_X^\top x_i$ and $W_Y^\top y_i$ subject to the constraints that the elements of the transformed vectors are pairwise uncorrelated and of unit variance. The optimization objective of CCA in matix form is given by:

$$\max_{W_X, W_Y} \quad tr(W_X^\top X Y^\top W_Y) \tag{1}$$

$$\text{subject to} \quad W_X^\top X X^\top W_X = W_Y^\top Y Y^\top W_Y = I \,. \tag{2}$$

After applying the projections W_X and W_Y, the vectors from both input spaces become d dimensional vectors. Denote these transformed vectors as $\tilde{x} = W_X x_j$ and $\tilde{y} = W_Y y_j$. Denote by $v(r)$ the r-th coordinate of a vector v. Eq. 1 is equivalent to maximizing the sum of the *canonical correlations* which are the d scalars defined as $\sum_{i=1}^n \tilde{x}(r)\tilde{y}(r)$, for $1 \le r \le d$. This is done subject to constraints of the form: $\sum_{i=1}^n \tilde{x}(r)\tilde{x}(s) = \sum_{i=1}^n \tilde{y}(r)\tilde{y}(s) = \delta_{rs}$, where δ_{rs} is defined to be 1 if $r = s$, and 0 otherwise.

A numerically favorable way to solve the CCA optimization problem is by means of Singular Value Decomposition (SVD) and matrix square root [GVL96], namely the decomposition of the matrix

$$M = (XX^\top)^{-\frac{1}{2}} XY^\top (YY^\top)^{-\frac{1}{2}} \,. \tag{3}$$

Let $M = USV^\top$ denote the Singular Value Decomposition of M, the canonical correlations are readily available as the diagonal elements of S. W_X and W_Y, which are not required by our technique, can be readily extracted from U and V.

In our application, CCA is computed over two types of vectors: a typically long view-based vector x_j of d_X elements, and a much shorter action-based vector y_j or length d_Y. The analysis is performed to short time windows of $n = 40$ frames or less, where each frame i constitutes one matching pair (x_i, y_i), and we set $d = n$. To directly compute the CCA as above, one would need to construct and invert matrices that have as many rows and columns as the length of x_i. Next, we describe how to transform the problem to a smaller one, which involves much smaller matrices, which are of the size of the time window. This technique can be seen as a simplification of kernel CCA [Aka01].

Let K_X, K_Y be the Gram matrices which are $n \times n$ matrices such that $(K_X)_{ij} = x_i^\top x_j$, and $(K_Y)_{ij} = y_i^\top y_j$. Straightforward linear algebra arguments imply that we can write $W_x = XA$ and $W_y = YB$ for two matrices A and B of size $n \times d$. Then, the correlation matrix (Eq. 1) is expressed as

$$W_x^\top X Y^\top W_y = A^\top X^\top X Y^\top Y B = A^\top K_X K_Y^\top B \;,$$

and similarly for the constraints (Eq. 2)

$$\begin{aligned} W_X^\top X X^\top W_X &= A^\top X^\top X X^\top X A &= A^\top K_X K_X A \\ W_Y^\top Y Y^\top W_Y &= B^\top Y^\top Y Y^\top Y B &= B^\top K_Y K_Y B. \end{aligned}$$

This yields a standard CCA problem where X and Y are replaced by the Gram matrices K_X and K_Y, this time solving for A and B:

$$\max_{A,B} \quad tr(A^\top K_X K_Y^\top B)$$

$$\text{subject to} \;\; A^\top K_X K_X^\top A = B^\top K_Y K_Y^\top B = I \,.$$

The proposed SVD solution still holds, only X and Y are replaced by K_X and K_Y respectively, and the decomposition is performed to matrices of size $n \times n$. Since n is typically small ($\tilde{4}0$) we are able to perform hundreds of such decompositions per second.

3.2. Measuring the correlation

At each point in time, each view stream k is represented by a set of vectors $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$ which encode the last n frames as seen at this particular view. We employ the simplest encoding and use the low resolution rendering of the scene as the representation of each view. This has the advantage of allowing faster rendering of views by conventional graphics cards.

The human motion parameters are represented by conventional means. At each point in time, the poses in the last n frames are encoded as a series of vectors $y_1, y_2, \ldots, y_n$. Within each such vector y_j, m joints are represented as $3m$ dimensions. Each single joint is encoded as three entries storing either the joint global location and rotation, or the skeleton relative location and relative rotation.

We employ CCA to the pairs $\{(x_i^{(k)}, y_i)\}_{i=1}^n$ arising from each view k to prioritize the expressiveness of views. If a view is informative regarding the underlying motion parameters, then we can expect that the correlation between the pairs of vectors to be high at least along one direction. Moreover, if the view captures several uncorrelated aspects of the underlying motion, we can expect to find significant correlation in several orthogonal directions. Therefore, for each view k we employ two statistics obtained by CCA: the first canonical correlation (h_k), and the number of canonical correlations above a certain threshold (n_k). The first measure expresses the dominance of the most prominent component which is shared by both the view and captured motion. The second measure expresses the number of uncorrelated directions in which the view and the scene motion are similar. The threshold ensures that the counted directions are strongly expressed in both the view and the parameterized motion. An

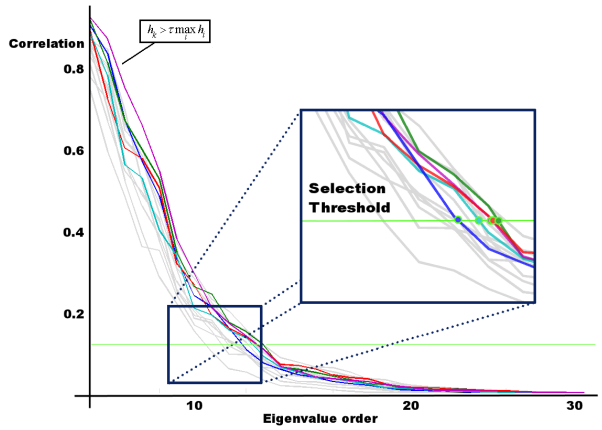


Figure 3: *The resulting canonical correlations for several views of a single character exercises scene. Here we highlight the curve characteristics used by our method - the largest canonical correlation h_k, and the number of canonical correlations n_k above a certain threshold.*

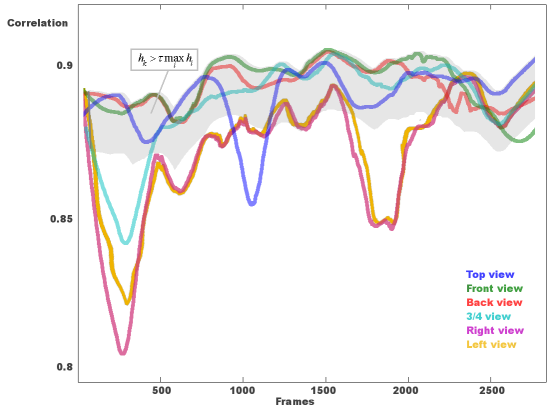


Figure 4: *The largest canonical correlation (h_k) as a measurement for the quality of a view . Results shown are of several views for a single character exercises scene. The gray ribbon indicates the views which will be considered during a shot switch, i.e., those with h_k that is sufficiently close to the highest one.*

example of the two measurements are shown in Figure 3. An example of the behavior of h_k over time for a simple scene of a single character is shown in Figure 4.

4. Method description

The availability of the correlation-based measures suggests a simple process for generating an online clip of a given animation scene from multiple views. The resulting clip is generated by continuously selecting views which are best representatives of the animation:- the method selects the view with the best quality measure as the current view stream, and use it until its accumulated erosion signal exceeded a given threshold. After which it select a different view and continue in the same manner. The details of each of these steps are described next.

4.1. View Streams Settings

The determination of a camera control path is known to be a time consuming task since the underlying optimization combines many constraints. Instead of selecting a single "best quality" view we use simple methods to suggest a set of different camera paths. To explore the stability of our solution we have implemented several different strategies for generating the candidate camera paths, including randomly generated camera paths and cinematography based path heuristics. Random constructions include simple random selection of static camera vantage points, and paths which comprise of random straight line and simple arcs as the camera path. Following common cinematography guidelines, we used several path generation methods including character tracking camera at three-quarters view, an over-the-shoulder camera, various simple vantage point cameras positioned on the front, back,right, and left sides of the scene, dollying along the scene movement line, and a distant "director shot" camera. An example of the large variety of selected paths is illustrated in Figure 7(b) and in Figure 9. While the differences in results of the various strategies will be discussed in Section 5, the following next steps were applied in the same manner to every set of camera paths.

4.2. Erosion of the current view

Cinematography suggests that views are switched in cases where the current view is no longer effective, when a better expressive view can be selected or when sufficient time had elapsed since the last switch [Ari76, HCS96]. To accommodate these guidelines, we use an accumulated view erosion signal (AVES). This is a monotonically increasing signal; Once it reaches a value of 1 a new view is selected.

AVES incorporates a time erosion, which is controlled by a switching time parameter $\bar{t}$ set to 90 frames (three seconds), the number of frames since the last view switch t, and the quality of the various views $h_k(t), k = 1..N$ in comparison to the quality of the currently active view $h_c(t)$. Specifically, $AVES(0) = 0$ and the accumulation formula is:

$$AVES(t) = AVES(t-1) + \left(1 - \frac{h_c(t)}{\max_k h_k(t)}\right) \cdot \frac{t}{\bar{t}} + \frac{1}{\bar{t}} \quad .$$

This formula limits the number of frames between view switch events to $\bar{t}$. In case where the highest quality view is not the active view, AVES promotes an earlier view switch. The progress of AVES depends on the relative qualities of the currently best view and the active view, as well as on the

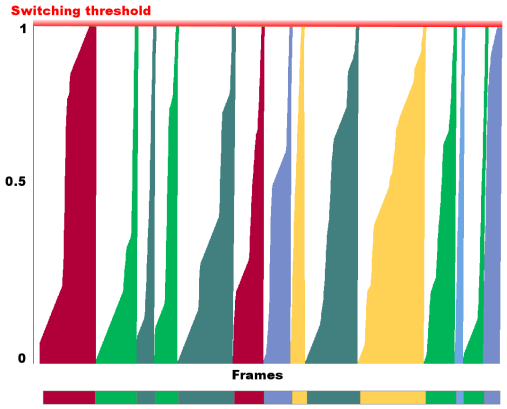


Figure 5: *The accumulated view erosion signal (AVES) function. When the signal exceeds a certain threshold, a view switch is performed.*

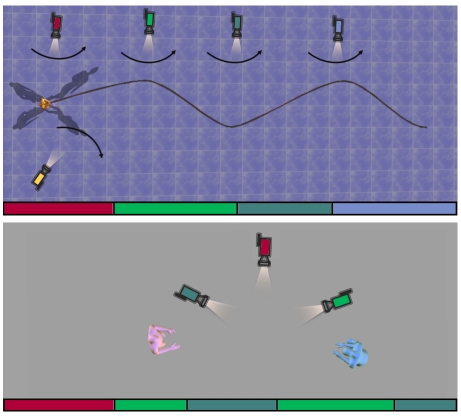


Figure 6: *Our method shot selections mimics the behavior defined in known cinematic idioms. The top row shows a setup which captures a long character walk. Bottom row: the camera setup for a conversation-like scene. Each of the images show the resulting camera shot sequence, color coded according to the used cameras. For better impression we refer the viewer to the accompanying view to this paper.*

amount of time since the last view switch. The latter ensures that shots do not change too rapidly. Figure 5 illustrates the accumulation term behavior over time.

4.3. Switching to a new view

In considering which view would become the next view, the view quality score h_k plays an important role in conjunction with geometric considerations, which eliminate views that will make the video clip incoherent. First, to avoid what is known as "jump cut" the spatial lookat angle difference between the previous and the new view is constrained to be above a given threshold. Second, in fast action movements the new and old views are constrained to be on the same side of the scene progress line. Both of these constraints are easily calculated, and geometrically unsuitable views are dismissed early to avoid unnecessary computations.

Further, we rule out the views which have relatively low correlation with the animation, as depicted by the first canonical correlation h_k (Sec. 3.2). Specifically, we only consider candidate views k for which $h_k > \tau \max_i h_i$, for some parameter $\tau = 0.9$.

Finally, we evaluate the candidate views by examining a combination of the number (n_k) of canonical correlations above a threshold of σ and the number of character face pixels. The latter is estimated rapidly as the area of the projection of a disk placed on top of the character's face onto the image. In our system $\sigma = 0.1$ and the score is computed as $q_k = n_k + f_k$. The first term, ensures that the selected view incorporates sufficient motion components, and the second biases the quality toward ones with frontal character views.

5. Results and Discussions

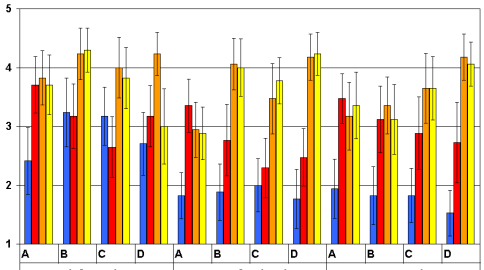
5.1. Cinamatic idioms

cinematographic idioms include guidelines for camera work for specific scenes, which describe a preset cameras placement, and a switching policy between them [Ari76, Mas65]. Using these idiom-based preset camera locations for our method cameras, we can observe interesting similarities between our method selection policy and the idioms switching guidelines, as shown in Figure 6.

Single character action. We have placed cameras at different directions around a exercising character (top, bottom, sides, frontal and back views and three-quarter views). While examining the output of our method, it becomes apparent that for a single-character the highest quality (h_k) views are often the three-quarters view, and in several cases, the top view. Incorporating the face area cue, the three-quarters view becomes significantly superior to the top-view and all other standard views. This result matches the known cinematic rule in which character actions should be captured from the three-quarters view, since it best represents the depth of character while performing the actions [Ari76].

Motion along a path. Capturing a character walk along a path, we placed cameras at uniform intervals. The method generates a sequence of long panning shots. This behavior is the suggested idiom for capturing large movement of characters over a vast scenery [Ari76].

Back and forth interaction between two persons. Back and forth interaction is common, e.g., in dialogue scenes. Here, we've placed three cameras, as suggested by similar idioms - two cameras were placed pointing to the figures,



	A	B	C	D	A	B	C	D	A	B	C	D
	informative				professional				expressive			
Naive	2.41	3.24	3.18	2.71	1.82	1.88	2.00	1.76	1.94	1.82	1.82	1.53
Our	3.71	3.18	2.65	3.18	3.35	2.76	2.29	2.47	3.47	3.12	2.88	2.72
Animator[1]	3.82	4.24	4.00	4.24	2.94	4.06	3.47	4.18	3.18	3.35	3.65	4.18
Animator[2]	3.71	4.29	3.82	3.00	2.88	4.00	3.78	4.24	3.35	3.12	3.65	4.06

Figure 8: *Results of our user study comparison. Participants were shown 4 scenes, each rendered by 4 camera control methods (shown in different colors). They were asked to grade the information,professionalism and expressiveness of the clips from 1-5 (5 being the best score).*

and a third was placed to capture both. Due to the lack of audio or lip-movement data, we used a ball passing sequence, which is analog in this respect to a verbal dialogue. The selected camera sequence significantly resembles the sequence suggested for shooting a conversation between two persons, according to their "turn" in the dialogue [Ari76].

5.2. Evaluation

We evaluate the results of our method, by using several means. First we compare it to the results of the recent work of Assa *et al.* [ACO*08]. As their work only handles scenes with limited complexity and single characters, we further continue with an evaluation of our results by performing a user study, similar to their study, which compares our results to that of a professional animator and a naïve camera control method.

The numerical comparison is performed using the view metrics suggested by [ACO*08]: (1) the total energy E of the generated camera path; (2) the external energy term, $E_{external}$ which includes the saliency biased spatial viewpoint potential (without the camera smoothness term); and (3) the "standard viewpoint potential metric" V. Note that as stated by [ACO*08] the three metrics measure only specific aspects of the animation, and the professional human animator obtains worse results in those metrics than the method of [ACO*08]. In all three measures, our method energy is between their reported results of a human animator and their method values. The difference between the approaches is best illustrated in Figure 7, where we show the camera path of Assa *et al.*, an example of our random views, and the resulting selected views. For a better impression and comparison we direct the reader to examine the supplementary video material of this work.

Following the evaluation methods used in recent studies, we compared the information, professionalism and expressiveness on several clips. The compared methods include the most correlative single camera location (*Naive*), our method (*Our*), and two professional animator results - with no design restriction (*Animator*[1]), and with camera switching every 3-4 seconds restriction (*Animator*[2]). For a better impression we direct the reader to examine the supplementary video material of this work. The clips were presented in a random order to 20 computer graphics students, who were asked to grade from 1-5, 5 being the best score, according to how informative is the result, how professional it looks, and its expressiveness. The results from the user study, (Figure 8) indicate that our solution scores better then the single camera location option, and in some cases it matches the professional animator options score. In all three criteria a t-test indicates that our method significantly outperforms (p-value $< 5\%$) the best single location camera. The user study is presented in the supplementary material of this paper.

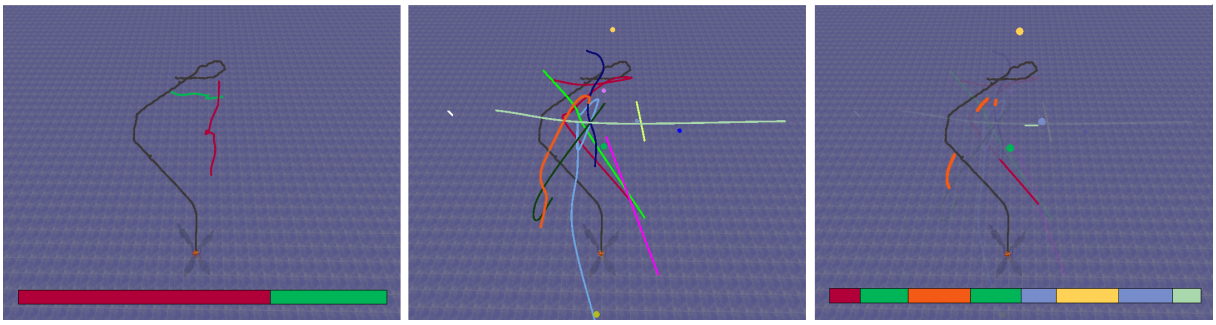
As part of the evaluation we have tested our system on cluttered scenes. Our method was able to deal with low to medium level of clutter. However, for very cluttered scenes, a view was sometimes picked in which the main action is occluded in the middle of the frame while the sides of the frame are uncluttered. This behavior differs from the typical choice of human directors. Note that all prior computational methods assume no clutter at all, and are not robust even to low levels of clutter.

As the method relies on viewpoint images, we examined its sensitivity to different character models, geometries and textures. When examining characters with significantly different proportions, as the ones shown in Figure 9 4th row, unorthodox viewpoints were consistently selected by our technique, such as the top and back views. These selections suggest that the currently common viewpoint selection convention is specifically tuned toward a standard human character proportions. Comparing different character textures, did not show any significant difference, as long as the character was sufficiently represented by the low resolution views.

We have evaluated the performance of our method, by using a variable number of cameras, starting from 3 up to 50 cameras in various scenes. As described earlier, we also used different camera placement strategies which include both random placement and movement, but also idiom-based configurations. Although such a comparison highly depends on the viewed motion, in many cases a small number (up to ten cameras) of the randomly selected viewpoints were sufficient to generate satisfactory results. Moreover, when using simple heuristics as the ones described in idioms configurations, usually a 4-5 set of cameras were sufficient to capture the motion in a pleasing manner.

5.3. Implementation details

The presented method does not include any computational bottlenecks. The various views are rendered at thumbnails



(a) (b) (c)

Figure 7: *A top level view of the generated character root (shown in dark gray), and the camera path used to capture the motion (color coded according to cameras). (a) illustrates the resulting camera shots of Assa et al., (b) illustrates the various camera movement path used by our algorithm (c) shows the automatically selected shot sequence. The sequence of color coded camera shots are shown below (a) and (c).*

of 64×48 pixels, and the CCA calculations are performed on 40×40 matrices, which are created by multiplying tall thin matrices by their transpose. All of the example clips shown in the accompanying video were generated at real-time speeds on a Pentium M 2.1GHz, with 1.5GB of memory, with standard ATI Mobility FireGL graphics card, and by using non-optimized Matlab and c++ implementations.

Our method is robust with regard to view resolution and quality. The thumbnail views we employ were found experimentally to have the same level of performance as higher quality views, hence we are able to render very rapidly. The openGL based rendering of the views thumbnails were performed at 720 frames per second, which allows comparing up to around 20 distinct views, each rendered at 30 FPS. The fast CCA computation was performed at a pace of 100 CCA calculations per second, using Matlab. As a result, such a mid- to low-range system could support a realtime rate of about 20 views at 30 FPS. For scenarios such as computer games where the system resources are shared with other threads, we experimented with reducing the frame rate to 20 and 15 FPS, with minimal to no impact on the quality of the results. Additional reduction in the number of frames per second used for analysis resulted in deteriorating performance for fast moving action scenes.

We have tested the technique on various clips with variable number of potential views, by using complex and simple scenes, and multiple characters. Some of the results are presented in Figure 9 and in the accompanying results and user study videos.

5.4. Selection without animation information

We adopt the method to cases where only the view streams are provided. This problem is ill-posed since it is not clear which apparent motion is the significant one without semantic reasoning or additional cues. However, it is possible to

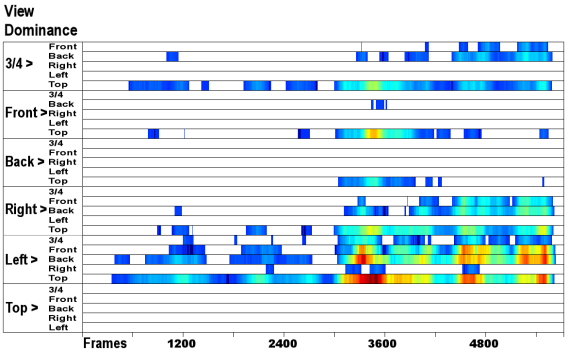


Figure 10: *Using multivariate linear regression to discover a partial order of views in respect to the motion they represent. This figure describes the results of the single character exercises clip, with 6 different views - top, front, back, right, left and 3/4 view. Each row presents the dominance of one view over the other (high dominance is shown in yellow, and no apparent dominance in white).*

infer relative dominance of some of the views. For example, comparing the relative quality of views depicting the same part of the scene may be possible, while there is no comparison between views of different scene parts.

Computationally, given a set of views of a scene, we examine the correlation of each view pair, by using *multivariate linear regression*. Given two views $x_i^{(j)}$ and $x_i^{(k)}$, $i = 1..n$, we obtain a matrix A_{jk} which minimizes $E_{jk} = \sum_i ||A_{jk}x_i^{(j)} - x_i^{(k)}||^2$, and a matrix A_{kj} that minimizes $E_{kj} = \sum_i ||A_{kj}x_i^{(k)} - x_i^{(j)}||^2$.

The resulting error estimate E_{jk} evaluates, using a linear approximation, how well view j captures the information of view k. This is run for time windows of $n = 100$

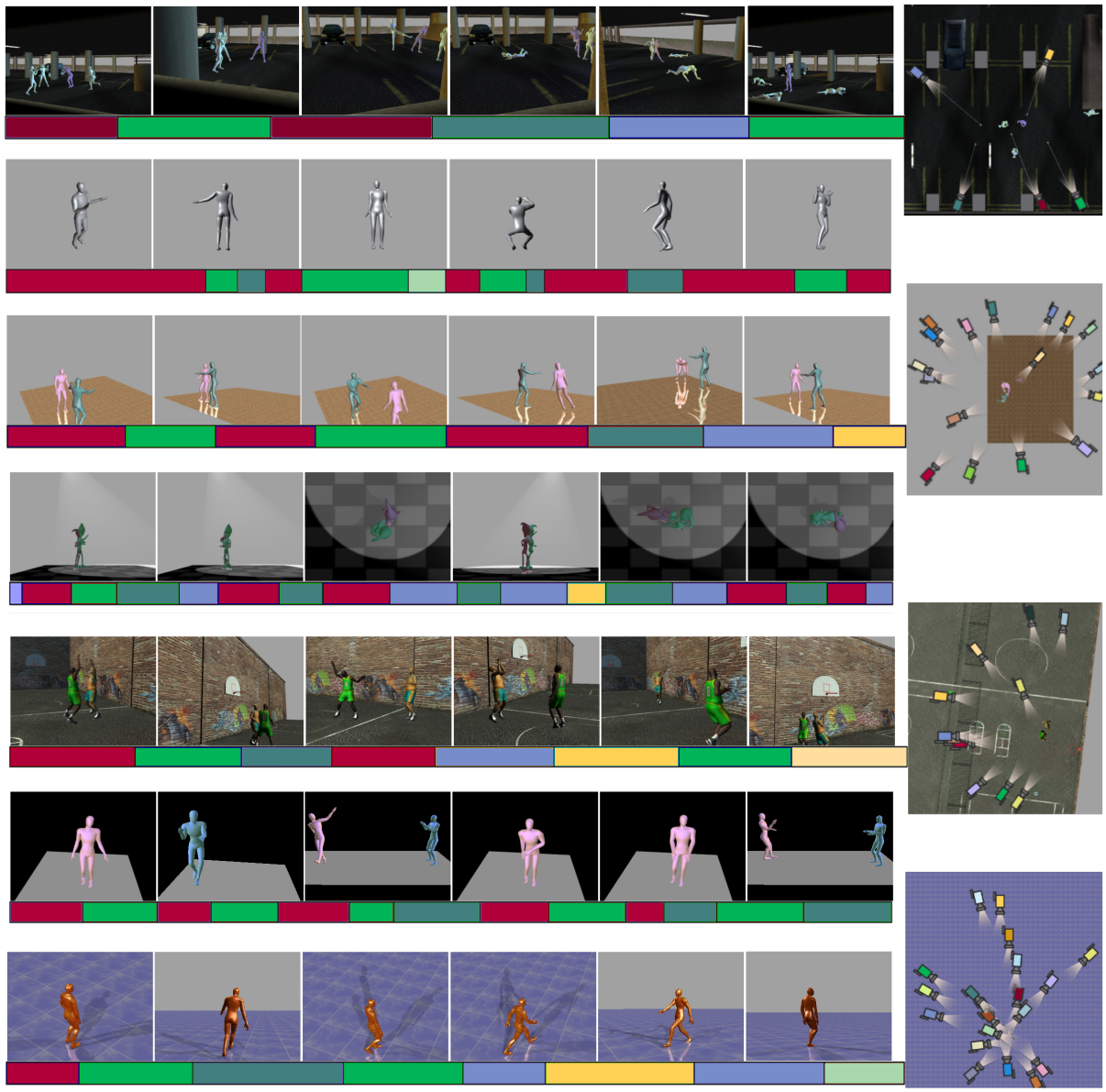


Figure 9: *An illustration of some of our results. Each of the 7 clips show 6 equally spaced view frames from the resulting clip, and the computed shot sequences, color-coded according to used cameras. The presented clips include (top to bottom), a fight scene, single character exercises, blindman's buff game, dance, basketball match, ball passing and a long walk. The color-coded camera initial configurations for clips 1,3,5,7 are illustrated on the right. We show here only a sample of the camera path selection strategies which were used. For a better impression of our method's results, we refer the reader to examine the accompanying video clip to this paper.*

frames. If $E_{jk} << E_{kj}$, we determine that view j dominates view k, and thus presents the underlying motion better. We employ this partial order to obtain at each point in time the list of all views for which there is no dominating view. Figure 10 shows an example, and another example is provided in the accompanying video on a difficult scene from the CMU quality of life grand challenge (http://kitchen.cs.cmu.edu).

5.5. Limitations

The main limitations of the proposed method emanate from its simplicity. Since the automatic method does not employ sophisticated detection and recognition algorithms, and the CCA correlates the view to the global animation, the selected views might include close-ups and even extreme close-ups shots. This problem can be minimized with adoption of stronger supporting heuristics, which will be used to ensure that the selected view is sufficiently inclusive. The semantic meaning of the view can also influence the quality of the result. Theoretically, since the method is not biased toward certain body parts, views which do not focus on the significant limbs may be selected. Incorporating a similar mechanism as described in Assa *et al.* [ACO*08] would provide a bias toward the significant limbs according to the performed action.

An additional limitation emerges from the method incapability to predict the potential future correlation between the views and the animation. In some cases, immediately after a certain view is selected, its quality quickly deteriorates, and as a result a new view selection should be made. Systems which attempted to predict the scene progress, as the work of Halper *et al.* [HHS01] did not produce a significant improvement. A practical approach for solving this problem can be resolved by creating a short delay (a couple of seconds) between the inputs and the resulting stream, allowing a look ahead into the future and selecting accordingly.

6. Summary

In this work, we present a novel technique which is based on exploring the animation-view correlation, introducing a fast and simple method for multi-shot video overview of a given human-action animation. This technique is shown to fit several known cinematic idioms, as well as to produce satisfactory results compared to previous work in this field. The applications of techniques for real-time human motion aware camera control system are vast and include both autonomous camera system for gaming and virtual environments. Such methods also suggest tools to guide novice animators in building a preliminary video clips which highlight their animation. This can be extended to automatic generation of video summaries of animation which can be created on demand in an online manner.

Although we did not specifically present in this work how to use additional signals to correlate to the view and the motion, it is easy to foresee usage of similar methods which synchronize music, voice and other cues. This may result in additional methods which can promote automatic music video generation. The presented extension to the case where no animation model is provided holds potential for a low-overhead multi-camera live video blogging.

References

[ACCO05] ASSA J., CASPI Y., COHEN-OR D.: Action synopsis: Pose selection and illustration. In *SIGGRAPH 2005 Conference Proceedings* (Aug. 2005), vol. 24, ACM, pp. 667–676.

[ACO*08] ASSA J., , COHEN-OR D., , YEH I.-C., LEE T.-Y.: Motion overview of human action. *ACM Transactions on Graphics (Also Proceedings of ACM SIGGRAPH ASIA) 27*, 5 (2008).

[Aka01] AKAHO S.: A kernel method for canonical correlation analysis. In *Int. Meet. of Psychometric Soc.* (2001).

[Ari76] ARIJON D.: *Grammar of the film language*. Silman-James Press, 1976.

[Bor98] BORGA M.: *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, 1998. Dissertation No 531, ISBN 91-7219-202-X.

[BTM00] BARES W. H., THAINIMIT S., MCDERMOTT S.: A model for constraint-based camera planning. In *Smart Graphics, Papers from the 2000 AAAI Spring Symposium* (2000), vol. 4, AAAI Press, pp. 84–91.

[CMN*05] CHRISTIE M., MACHAP R., NORMAND J.-M., OLIVIER P., PICKERING J.: Virtual camera planning: A survey. In *Smart Graphics* (2005), pp. 40–52.

[CO06] CHRISTIE M., OLIVIER P.: Camera control in computer graphics. In *Eurographics 2006 Star Report* (2006), pp. 89–113.

[DZ94] DRUCKER S. M., ZELTZER D.: Intelligent camera control in a virtual environment. In *Proceedings of Graphics Interface '94* (1994), pp. 190–199.

[GVL96] GOLUB G. H., VAN LOAN C. F.: *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[GW92] GLEICHER M., WITKIN A.: Through-the-lens camera control. In *SIGGRAPH 1992 conference proceedings* (New York, NY, USA, 1992), ACM, pp. 331–340.

[HCS96] HE L.-W., COHEN M. F., SALESIN D. H.: The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *SIGGRAPH 1996 Conference Proceedings* (1996), ACM, pp. 217–224.

[HHS01] HALPER N., HELBING R., STROTHOTTE T.: A camera engine for computer games: Managing the trade-off between constraint satisfaction and frame coherence. In *EG 2001 Proceedings* (2001), vol. 20(3), Blackwell Publishing, pp. 174–183.

[HO00] HALPER N., OLIVIER P.: Camplan: A camera planning agent. In *AAAI 2000 Spring Symposium on Smart Graphics* (2000), AAAI Press, pp. 92–100.

[Kat91] KATZ S. D.: *Film Directing Shot by Shot: Visualizing from Concept to Screen*. Michael Wiese Productions, 1991.

[KL08] KWON J.-Y., LEE I.-K.: Detemination of camera parameters for character motions using motion area. *The Visual Computer 24* (2008), 475–483.

[LST04] LIN T.-C., SHIH Z.-C., TSAI Y.-T.: Cinematic camera control in 3d computer games. In *WSCG* (2004), pp. 289–296.

[Mas65] MASCELLI J. V.: *The Five C's of Cinematography: Motion Picture Filming Techniques.* Cine/Graphic Publications, 1965.

[MK06] MCCABE H., KNEAFSEY J.: A virtual cinematography system for first person shooter games. In *Proceedings of International Digital Games Conference* (2006), pp. 25–35.

[VBP*09] VIEIRA T., BORDIGNON A., PEIXOTO A., TAVARES G., LOPES H., VELHO L., LEWINER T.: Learning good views through intelligent galleries. In *Eurographics* (Munich, march 2009), pp. 717–726.