The Visual Computer manuscript No.
(will be inserted by the editor)

Yaron Caspi · Anat Axelrod · Yasuyuki Matsushita · Alon Gamliel

Dynamic Stills and Clip Trailers

Abstract We propose a method for generating visual summaries of video. It reduces browsing time, minimizes screen-space utilization, while preserving the crux of the video content and the sensation of motion. The outputs are images or short clips, denoted as *dynamic stills* or *clip trailers*, respectively. The method selects informative poses out of extracted video objects. Optimal rotations and transparency supports visualization of an increased number of poses, leading to concise activity visualization. Our method addresses previously avoided scenarios, e.g., activities occurring in one place, or scenes with non-static background. We demonstrate and evaluate the method for various types of videos.

Keywords Video Summaries · Key-Pose Selection

1 Introduction

Different types of media have associated promotion mechanisms. Books have their covers, and movies have their trailers. In both examples, promotion is coupled with content information. In this work, we develop a similar mechanism for an emerging type of medium - video clips. We develop means for efficiently extracting and visualizing the content of short video clips using two display methods: (i) *Dynamic still* - a single image conveying the activity by combining information from different temporal moments; (ii) *Video clip trailer* - a few seconds long video that displays the clip's highlights. We believe that such a mechanism will greatly enhance many domains where visual summaries are desired such as online video libraries or news web pages (e.g., [11,12,14]).

Current video libraries use text, a single image (usually the clip's first frame), or a set of key-frames for depicting the content of videos. Since the differentiating factor of videos from images is motion, we believe that the expressive power of text and key-frames is limited. Thus, our goal is embedding and visualizing dynamics in a concise representation, while maintaining the clarity and sensation of motion.

This work focuses on clips of human subjects. Video objects are extracted from different time-frames, and several poses that best represent the activity captured in the video are concurrently displayed. Technically, we treat a video clip as a three-dimensional (3D) volume with the spatial axes X, Y and temporal axis T. To reduce the inherent spatial redundancy in the video, we cut out the foreground objects from this volume [17,25] and compute alpha matting values [24]. We further reduce temporal redundancy of the foreground data by automatically selecting key-poses from the sequence of extracted objects, utilizing a new algorithm which takes shape information as input. This data reduction allows handling of previously neglected scenarios. Unlike synopsis mosaics [15] that only treat videos of objects translating from side to side, we also consider self-occluding activities (Figure 2). The resulting output from the above process is a set of key-poses selected from segmented objects. These poses with associated alpha masks are denoted as *pose slices* and can be rendered in two different ways. First, they can be composed into a single static image that we call a *dynamic still*. Second, they can form a few second long video presenting the essence of the activity. We term this representation a *clip trailer*.

We provide a set of manual and automatic tools that enable non-professional video editors to generate dynamic

Yaron Caspi
Faculty of Mathematics and Computer Science
The Weizmann Institute of Science
E-mail: yaron.caspi@weizmann.ac.il

Anat Axelrod
School of Computer Science
Tel Aviv University
E-mail: anataxel@post.tau.ac.il

Yasuyuki Matsushita
Visual Computing Group
Microsoft Research Asia
E-mail: yasumat@microsoft.com

Alon Gamliel
School of Computer Science
Tel Aviv University
E-mail: gamliela@post.tau.ac.il

Fig. 1 "Ascending and descending stairs". Visualizing an activity Eadweard Muybridge, 1887.

(a) Non self-occluding

(b) Self-occluding

Fig. 2 Self-occluding vs. non self-occluding activities. While synopsis mosaics work for non-self occluding activities (a), they face difficulties with self-occluding activities (b). Our approach handles both cases (see Figure 12).

stills and clip trailers. These include: rotation and zoom of the entire set of pose slices and/or each pose slice independently, and use of transparency for adding spatial or temporal context and for emphasizing particular activities. Our visual content summaries have a range of application areas, such as web sites where users download video clips or cellular phone download centers. Summaries might also be used as static or "breathing" thumbnails on a PC, where several clip trailers are played automatically in an icon size display. We illustrate these applications using a variety of clips from home and sports videos.

The rest of this paper is organized as follows. Section 2 reviews prior work. Section 3 discusses motivating observations. Section 4 reviews the technical details of our method. Results and applications are discussed in Sections 5 and 6. Section 7 concludes the paper.

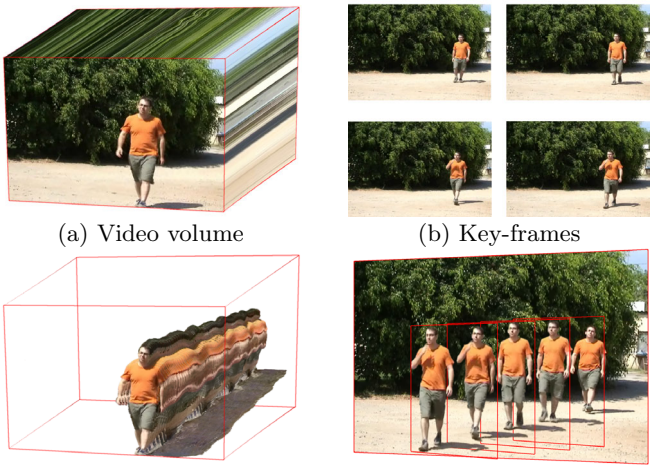
2 Related work

Representing motion and activity has been a challenging problem faced by many visual art masters and scientists. Photography entrepreneurs such as Muybridge, Marey and others developed multiple-exposure techniques for visualizing human activities (Figure 1). Advances in computer graphics and computer vision allowed spatially stitching several images together to create a "panorama" (e.g., [26]). Panoramas have been used for representing videos. For example, Taniguchi *et al.* [27] combined shot detection and panoramas to generate a mixed catalog of key frames and panoramas. Irani *et al.* [15] superimposed copies of foreground objects taken at different time points on a background panorama. To avoid overlaps between different copies, the video is diluted by sampling frames uniformly or manually [20]. Recently, an approach of stroboscopic visualization of movement based on interactive digital photo-montages was proposed by Agarwala *et al.* [2]. Similarly, Chiu *et al.* [10] suggested arranging blobs extracted from key-frames in a collage representation, as a way to improve screen utilization when visually summarizing business meetings. However, the above processes are only effective for non self-occluding activities, and fail when objects appear at the same image location more than once (Figure 2). Our framework handles this temporal self-occlusion problem.

Methods for visualizing dynamic scenes have been studied not only for producing still images such. Agarwala *et al.* [3] and Rav-Acha *et al.* [22] manipulated both space and time to generate a wide field-of-view panoramic video. Interactive artistic manipulations of such video volumes are implemented in the Khronos Projector [9]. Unlike this work, the above methods do not reduce the size of the represented data, and most of them even increase it for producing the new representation.

Since our focus is on video objects rather than video frames, we employ a video object cutout process. We have many options in choosing a video object cutout algorithm. Interactive approaches such as [17,28] are both useful for a cluttered background or a non-stationary camera. Alternatively, if the camera is static and the background is also static, recently proposed automatic methods [13,25] may be applied. Segmented objects were also used for the purpose of interactive video browsing [6]. However, standard image and video outputs are not supported there, and user interaction is required to select highly informative key-poses.

Our video processing pipeline involves an analysis of the pose motion for the purpose of key-pose selection. This problem has been studied in the context of video summarization and retrieval, traditionally achieved by key-frame selection to approximate the task of key-pose selection. Recently, key-pose selection algorithms that focus on video objects rather than video frames are gaining more attention. Loy *et al.* [19] computed a frame-to-frame distance based on contour matching, while Liu *et al.* [18] used distance information obtained from motion-capture data. Both methods cluster frames and select the cluster centers as key-frames. Assa *et al.* [4] suggested that a concise and informative synopsis of a human activity is better estimated by selecting the extreme poses of the motion. In this work, we follow Assa *et al.* 's approach of selecting extreme key-poses and extend it to 2D video sequences. While Assa *et al.* 's work mainly considered tracking of skeleton joints usually obtained from motion-capture data, we achieve a similar goal only from 2D foreground images.



(a) Video volume (b) Key-frames

(c) Object segmentation (d) Pose slices

Fig. 3 Different video visualizations. (a) A video visualized as space-time volume. (b) A key-frames representation. The representation is highly redundant and lacks the sensation of motion (c). A space-time volume of the moving object. This representation does not suffer from background redundancy, but still hides the activity details. Our dynamic still representation in (d) combines a single background frame with a set of key-poses to concisely depict the activity.

3 Approach

The main challenge in generating an informative visual summary is dealing with the huge redundancies inherent in video. This section discusses several fundamental observations regarding video clips. Based on these observations, we then briefly describe the proposed framework.
(1) Background has a huge temporal redundancy. This is illustrated by the complete video volume in Figure 3(a). This observation leads to today's approach of key-frame representation, in which temporal redundancy is reduced (Figure 3(b)).
(2) Foreground is more attractive and informative than the background. In Figure 3 the walking person is clearly the most important portion in the scene. Furthermore, the background contributes mostly as an atmosphere, and its details are not crucial for understanding the scene. Consequently, we cut the background off from the volume (Figure 3(c)), and place it once as the background of the final output.
(3) Not all poses have the same importance. Poses at the beginning and the end of each consecutive motion are more important than intermediate poses.
(4) Background and foreground are not very sensitive to small rotations from the standard "head-on view". Hence, we encourage manipulations of the viewing angle. A pleasing result is illustrated in Figure 3(d). Note that our display size is on the same order of a single video frame's. Accordingly, the essence of the video clip is depicted by decreasing the redundancy in both background and foreground, and by choosing important poses of foreground that we call pose slices.

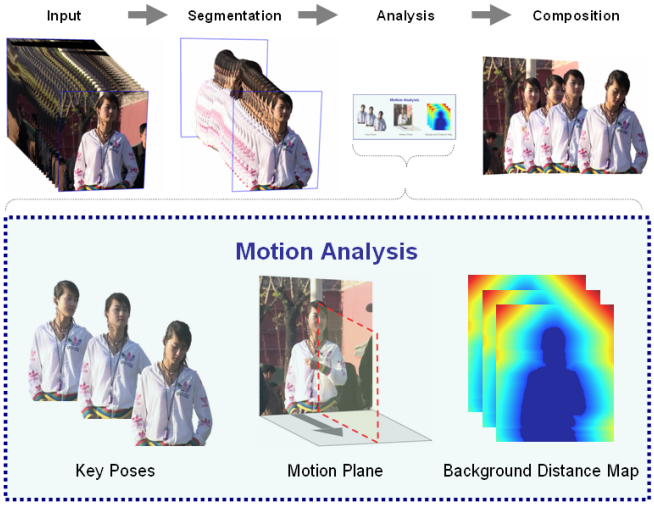


Fig. 4 Components of our framework. First, the input video volume is decomposed into foreground and background. Then, the motion of extracted objects is analyzed to produce key-poses and rendering instructions. These are used to compose the final visualization.

4 Technical Description

The proposed processing steps are illustrated in Figure 4. First, the input video volume is separated into foreground and background. Then, the motion and pose information of extracted foreground objects is analyzed, and highly informative key-poses are selected. Prior to composition, we compute an optimal viewing angle that minimizes the mutual occlusion of different poses. The composition step allows the author to define a set of rendering instructions to control the visualization effects. For producing dynamic stills, the visualization parameters such as the viewing angle and key-pose opacity are fixed. For a clip trailer, these parameters are time dependent. The next sections describe the major steps used in this process.

4.1 Video decomposition

The video decomposition begins with extracting foreground objects. Recently, we witnessed significant progress in this field. Li *et al.* [17] and Wang *et al.* [28] described interactive approaches, and showed that it is possible to segment complicated scenes with a user's guidance. Sun *et al.* [25] and Criminisi *et al.* [13] described methods of foreground extraction in real time for scenes with stationary background. Our method is not constrained to a particular video segmentation method, and the choice depends on the complexity of the scene. This is illustrated in Figure 5. For the simple stationary background in (a) the method of [25] was used, while for the cluttered scene in (b) the method of [17] was used. After video object segmentation, we apply an alpha matting

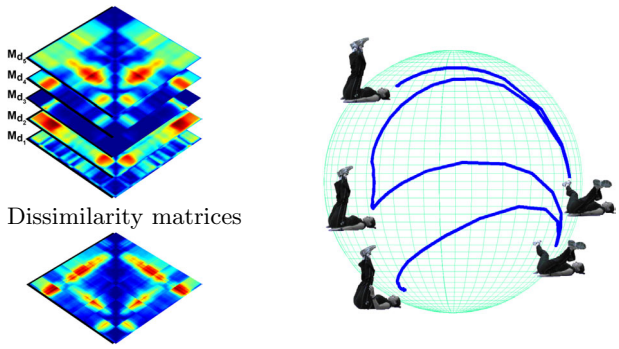
Fig. 5 Segmentation and scene complexity. The simple scene in (a) was segmented automatically in realtime, while the cluttered scene in (b) was segmented interactively. Final results are shown in (c).

algorithm, crucial for achieving a high-quality visualization. In our implementation, we use the Poisson matting algorithm [24].

4.2 Motion extremum analysis

This section describes an automatic method for selecting key-poses. The input is a sequence of segmented instances of foreground objects from all the frames in the video, denoted here as poses. This defines a curve in the huge space of all poses. We argue that important poses are extremum points on this curve, since activity is best illustrated by changes in motion. Such extremum points simplify the interpolation process for completing the activity representation. Thus, we call our approach *motion extremum analysis*. The output of this module is "pose slices". Figuratively, we slice the 3D blob of the foreground object's data and select poses with high importance for depicting activity.

Our method for motion extremum analysis is composed of the following steps: (i) Computing several pose-to-pose dissimilarity measures organized in dissimilarity matrices. These matrices are denoted M_{d_i}, where d_i is the i-th dissimilarity measure. Each matrix is of size $F \times F$ where F is the number of poses extracted in the segmentation step (Figure 6(a)). (ii) Combining these measures to produce an $F \times F$ unified dissimilarity matrix (Figure 6(b)). (iii) Embedding poses on a low-dimensional sphere, forming a *motion curve* (Figure 6(c)). The analysis of this motion curve enables the evaluation of the importance of poses. The correspondence between extremum poses and motion curve cusps is illustrated in Figure 6(c).



(a) Dissimilarity matrices (b) Unified matrix (c) Motion curve

Fig. 6 Motion extremum analysis. Several dissimilarity matrices (a) are merged into a single unified matrix (b). Poses are then embedded into a low-dimensional space. The cusps of the resulting motion curve (c) correspond to extremum poses of the activity (wide-open/closed legs).

Inter-pose dissimilarity measures. Pose-to-pose dissimilarity is measured using features computed for each pair of foreground silhouettes: (i) Orientation of the major axis, (ii) Elongation, (iii) Solidity, (iv) Shape overlap, and (v) Centroid position. Other features were tested, e.g., the velocity of the silhouettes' centroids, and the radii of the silhouettes' circumcircles centered at the centroids. However, these were discarded due to the small weights they were given during the unification process (see below).

(i) Orientation of the silhouette's main axis is computed using Principal Components Analysis (PCA) on the set S of 2D pixel coordinates inside the silhouette. The orientation dissimilarity $M_{d_1}(i,j)$ between the poses in frames i and j can be defined as:

$$M_{d_1}(i,j) = 1 - |\cos \alpha_{ij}| \,,$$

where α_{ij} is the angle between the main axis of pose i and that of pose j.

(ii) Elongation r_i of pose i, is measured by the ratio of the pixel distribution's variance along its primary and secondary axes in the eigen-space. It is measured by the first and second eigenvalues of the silhouette covariance matrix. This property differentiates thin elongated shapes from circular shapes. The elongation dissimilarity between poses i and j can be defined as:

$$M_{d_2}(i,j) = |\log(r_i/r_j)| \,.$$

(iii) The solidity of a foreground object is defined as the ratio between the area inside the silhouette and the area of its bounding box. This measure is useful for detecting the extension of limbs, the waving of hair or clothes, etc. The solidity dissimilarity $M_{d_3}(i,j)$ is defined similarly to $M_{d_2}(i,j)$ by substituting r_i and r_j with solidity values.

(iv) Shape overlap is computed using the absolute correlation [8] between pairs of registered silhouettes. Registration is achieved by aligning the silhouettes' centroids. The absolute correlation is minimized over a small search

radius R to reduce the sensitivity to subtle segmentation errors:

$$M_{d_4}(i,j) = \min_{dx,dy} \sum_{(x,y)\in B} |S_i(x+dx, y+dy) - S_j(x,y)|,$$

where B is the bounding box surrounding both silhouettes.
(v) The position dissimilarity is defined as the Euclidean distance between silhouettes' centroids. A few other measures (velocity, radius and area) were also tested. These, however, did not contribute to the unified representation of all dissimilarity measures (next paragraph), and therefore have been omitted.

Unified dissimilarity matrix. A unified dissimilarity matrix M is computed, aggregating the different inter-pose dissimilarity matrices M_{d_i}. We start by converting each dissimilarity matrix into a corresponding cross-product matrix P_{d_i}, as in multidimensional scaling (MDS), and normalizing each matrix by its first eigenvalue. An aggregation matrix is computed as a linear combination of the normalized matrices, and is then converted back into the unified dissimilarity matrix M. The weights for the linear combination are derived from a non-centered PCA of the P_{d_i} matrices [1]. This gives an "optimal" solution in the sense that measures which agree the most with each other contribute the most to the unified dissimilarity matrix.

A low-dimensional motion curve. Based on the unified dissimilarity matrix, we embed the sequence of input poses into a low-dimensional space. For this task, we use the spectral clustering method [21], a method designed for segmentation. It applies eigen-decomposition to a normalized version of a dissimilarity matrix, denoted as an affinity matrix. In our case, the affinity between poses is derived from their dissimilarity, as computed in the previous step. A spectral clustering method is applied to the temporally ordered sequence of points (poses). To incorporate this additional information and to avoid the influence of temporally distant poses, we multiply the entries of the affinity matrix $M(i,j)$ by an exponential decay factor $e^{-|i-j|/\delta}$, where δ is a constant which indicates that human motion segments are usually a few seconds long, e.g. 3 seconds. The output of the eigen-decomposition is a set of column eigenvectors ordered by their eigenvalues. The first k eigenvectors define an $F \times k$ matrix, where each pose i is associated with the i^{th} row of this matrix. In theory, spectral gap may be used to select k. In practice, a significant gap rarely exists. Therefore, global measures such as *stress* [16] should be used to select k. This set of F row-vectors are normalized to unit length and placed on the S^k unit sphere [21]. The advantage of this process is that it attempts to preserve local distance, thus similar poses are represented by adjacent points on the sphere. The advantages and relations of this approach to other embedding techniques (e.g., [23]) are discussed in [7].

Fig. 7 A dynamic still showing key-poses selected from a 2D animated sequence.

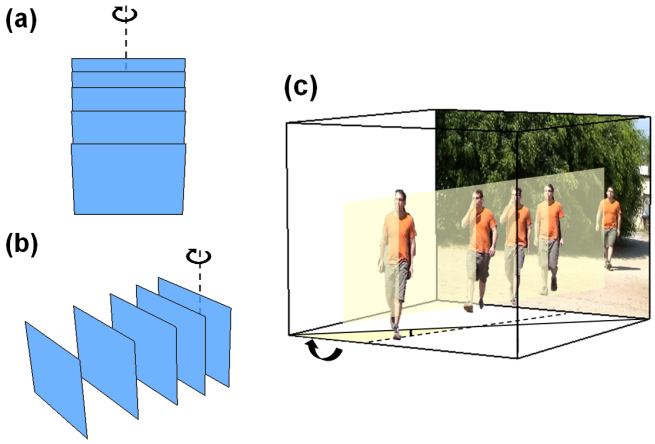


Fig. 8 Global rotation (a) and (b) is used to address occlusions between poses. (c) The optimal viewing angle is recovered by aligning the global motion-plane (yellow) with the major diagonal of the video volume, as indicated by an arrow.

Pose slicing. To select key-poses, we follow the approach of Assa *et al.* [4]. We iteratively choose poses which are locally extreme (distant from their temporal neighborhood). This is illustrated in Figure 6(c), where the sharp corners of the motion curve are selected and indeed correspond to extremum poses of the original motion. We compared our approach with Assa *et al.* 's method by rendering one of their animation sequences (Figure 7). Although our method does not take the explicit 3D motion data as input, it is able to generate a virtually equivalent result only from 2D silhouette data. This completes the video decomposition process.

4.3 Display composition

Once pose slices are selected, they are optimally visualized using their shape information and positions in the space-time video volume. This section introduces our approach for efficiently visualizing multiple pose slices in a limited screen space.

Global viewing angle. When visualizing multiple pose slices at the same time, self-occlusion of poses becomes one of the major problems. In our approach, self-occlusion of the pose slices is minimized by rotating the space-time video volume from the head-on view to a new viewing

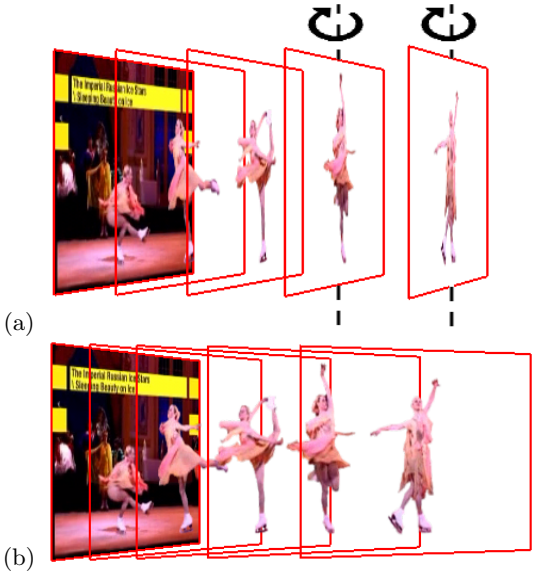


Fig. 9 After global rotation (a) some pose slices might appear flat. This is resolved by an automatic local rotation (b).

angle (Figure 8(a) and (b)). This rotation approach has an advantage; the orientation of the background plane gives a visual cue as to the direction of the time axis. An optimal viewing angle is computed automatically according to the global direction of the motion observed in the video volume. It can be computed by collecting 3D positions of foreground pixels throughout the video volume and fitting a 2D plane that minimizes the sum of perpendicular distances from the points. We implement this approach using PCA. The resulting 2D plane is denoted the *global motion plane*.

To minimize the deviation from the original viewing angle, we align the global motion plane with the closest diagonal of the 3D video volume. Figure 8(c) illustrates the global motion plane, and the recovered global rotation (marked in yellow). Note that this rotation depends on the spacing between poses. Namely, stretching the time axis reduces inter-occlusion and results in a smaller rotation.

Local viewing angle. The global viewing angle is optimized to provide the most informative visualization. However, a global rotation of the space-time volume results in a non-uniform viewing angle for different poses, and may cause some pose slices to appear flat (Figure 9(a)). A secondary rotation, we denote as the *local viewing angle*, is used to compensate for this flattening artifact. Our approach is motivated by visualization methods that use a cylindrical (or spherical) coordinate system (e.g., [26]). In contrast to a planar display, a cylindrical surface gives an orthogonal projection of surface points to a camera. At the same time, we wish to place poses along the Z-axis, in order to maintain relative positions of key-poses. Therefore, we only rotate each pose slice about its central axis to face toward the camera. In this way, the

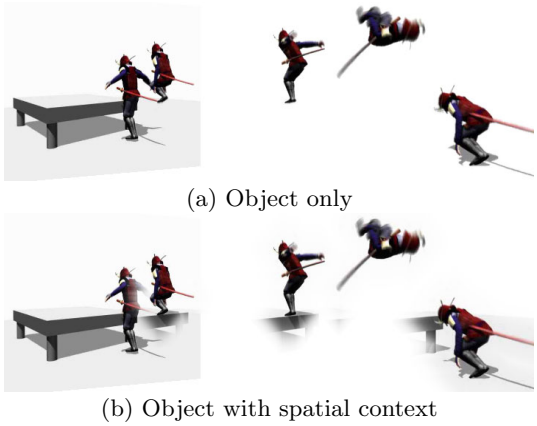


Fig. 10 Spatial context. (a) Only pose slices are visible; (b) shows the table that the samurai jumped from.

secondary rotation angle is automatically computed and applied (Figure 9 (b)).

Spatial context. In some video clips, a better understanding of the portrayed activity can be achieved by incorporating *spatial context* into the visualization. This is accomplished by visualizing additional "secondary" objects that appear next to the foreground object. These secondary objects can be extracted in the video object segmentation phase. Alternatively, image portions around the foreground object can be visualized by manipulating alpha values to create a gradual fade-out effect (Figure 10). In our case, alpha values are assigned using a sigmoidal function of the city-block distance d from the foreground object's boundary:

$$\alpha(d) = \frac{1}{1 + e^{\frac{d-\eta}{\sigma}}},$$

where η determines the width of the spatial context, and σ controls the strength of the fade-out effect. η and σ can be tuned interactively.

Temporal context. Another option to enrich the visual experience is to add pose slices. This, however, might cause a self-occlusion problem. Therefore, we add the

Fig. 11 Temporal context. The transparency of additional poses is based on their importance. Most informative poses (i.e., motion extreme points) are completely opaque.

(a) Key frames (b) Synopsis mosaic (c) Our representation

Fig. 12 Comparison of different video summaries. We compare the results of our visualization with previous methods using three different examples. The top row illustrates that our method can address more elegantly non static background, and a scene with multiple objects. The middle row illustrates how our method depicts a self-occluding activity. The third row compares summaries of a long home video.

poses (a) selectively using the key-pose selection results; (b) using transparency. Figure 11 displays an example of the temporal context effect. The number of poses to be added can be controlled interactively.

5 Discussion and evaluation

5.1 Benefits

The advantages of using our visualization are evident from the comparison of different methods shown in Figure 12. It is clear that our method captures the sensation of motion that is lost in a key-frames representation. Due to its better screen utilization, it provides a comprehensible visualization that allows a quick acquaintance with an efficient orientation in video sequences. It is also evident that previously proposed methods (i.e., synopsis mosaics) cannot address the above scenarios.

In the presence of large camera motions, our method should be applied after registration of the video sequence. The resulting space-time video volume is not a rectangular box. However, all the previously described steps apply in such cases as well. It is also interesting to note that a regular superposition of objects (e.g., [2,15]) is a special case of our method. Namely, when the object translates from side to side, the global rotation is effectively 0°, and the resulting dynamic still is essentially a dynamic panorama.

5.2 Limitations

The effective range of clips that our current implementation can address depends on the number of poses needed in order to describe a given activity. From our experience, the proposed visualization allows for about a dozen poses to be viewed simultaneously (i.e, in the same screen-space as a single frame). Obviously, our visualization can benefit from panoramic display and can depict longer activities. Other packing problems exist in the presence of multiple objects. While multiple objects may be displayed (see for example, Figure 12(c) top), there may be cases where rotations of distinct objects will occlude each other. Finally, the trade-off for the better screen utilization is that the exact spatial information of an activity is lost. The global rotation slightly shifts the locations of pose slices. An example where this may cause confusion is shown in Figure 13; although the girl was walking in place, it is perceived as forward motion.

5.3 User Studies

To verify the contribution of our representations we ran two user studies. One user study measured whether the clip trailers deliver accurate content information in a shorter time, and the other evaluates whether they were attractive.

Fig. 13 This dynamic still illustrates the limitations of depth ordering. While the person in the figure is walking in place, it is perceived as if she is moving forward.

Evaluating informativeness. Thirty users were asked to find unique events in video sequences or in the respective summaries. They are also asked to answer simple questions about them (e.g., describe the boy's emotions, who failed in returning a tennis serve). For each of the given 6 tasks, we tested the correctness of the answers and measured the time needed to complete the task. The results indicate that: (i) The fraction of correct answers was similar using both visualizations; (ii) The response time using the proposed video summaries was a few seconds (less than five), while using regular videos it had a linear dependency in the temporal length of the video clip, and was generally longer; (iii) Training with a single video summary significantly improved the success rate.

Evaluating attractiveness. One hundred and forty users were asked to select the most attractive video clip out of six options. Three of these were visualized by clip trailers and the other three were still images extracted from the videos. To overcome the inherent bias to particular video clips (e.g., clips that consisted of attractive objects), the selection of clip trailer or image for each video was reversed for half of the participants. The results indicate that in five out of six cases clip trailers were more attractive. Further details may be found in [5].

6 Applications and results

This section shows several results of our visual video summaries and illustrates the applicability of our video previewing techniques for different applications.

Video clips promotion. Clip trailers contribute to promotion of video clips. Therefore, a major application for clip trailers is in the domain of video media centers, such as video download web sites, Video-on-Demand interactive TV channels, or even video information and entertainment services featured on cellular phones. In Figure 12, we compare a key-frames representation of video clips as shown on a commercial web site with our innovative approach.

Breathing thumbnails. Another straightforward application of clip trailer technology is *breathing thumbnails*. Breathing thumbnails stands for a dynamic visual representation of home-users' PC folders containing video clips. This representation is built by concurrently displaying miniaturized clip trailers summarizing the video clips contained in the folder. It can be constructed manually or automatically if video segmentations are provided. An example of breathing thumbnails is illustrated in Figure 14.

Informative time line and indexing. Locating a specific clip from a clip archive is a difficult task, especially for videos with long and rich contents. One needs to tediously use the rewind/fast-forward buttons in order to understand what the clip is all about. On the other hand, clip trailers can show the major events in a video clip in a very informative way. Furthermore, our motion analysis that captures the key-poses may be used to generate an informative timeline. As illustrated in Figure 15, an informative timeline is constructed by rotating the pose slices to a 90° view angle, with the secondary rotation making them all nicely visible[1]. This technique, com-

[1] The data was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

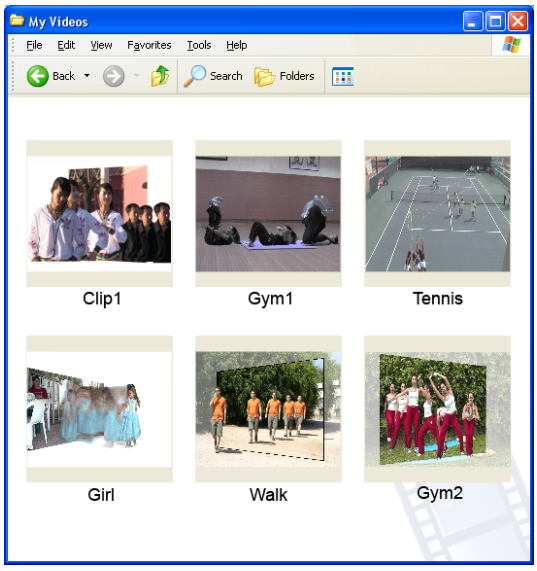


Fig. 14 Breathing Thumbnails. Iconized clip trailers can replace conventional static thumbnails. Such moving ("breathing") thumbnails are more informative than the clip's first frame that is currently used on regular PC's.

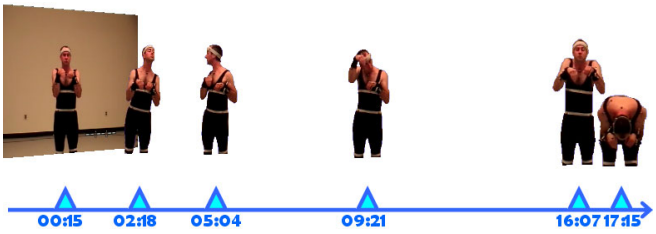


Fig. 15 Informative timeline.

bined with breathing thumbnails, simplifies the browsing process in a video archive.

7 Summary and future work

We have described means for an effective visualization of video clips. Our visualizations, dynamic stills and clip trailers, provide a better utilization of the display space, and therefore also a better utilization of time. Furthermore, our framework allows us to address scenarios that so far have been neglected (e.g., the presence of self-occlusion). The effectiveness of our approach is illustrated qualitatively by a large number of examples. It is also evaluated quantitatively with two user studies.

In the future, we plan to address some derived issues. First, video object segmentation remains the most time consuming part of our system. We would like to speed up this part, by combining proximity relations between coherent segments. Second, we wish to combine key-pose selection with occlusion constraints. By coupling the key-pose selection process with the resulting visualization, the screen space to information ratio and visual quality may be improved. This could be achieved by enforcing a new constraint on the key-pose selection algorithm.

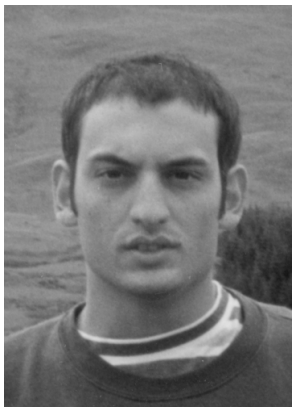
References

1. Abdi, H., Valentin, D., O'Toole, A.J., Edelman, B.: Distatis: The analysis of multiple distance matrices. In: EEMCV Workshop, pp. 42–47. San Diego, CA, USA (2005)
2. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. ACM Transactions on Graphics (SIGGRAPH) **23**(3), 294–302 (2004)
3. Agarwala, A., Zheng, C., Pal, C., Agrawala, M., Cohen, M., Curless, B., Salesin, D., Szeliski, R.: Panoramic video textures. ACM Transactions on Graphics (SIGGRAPH) **24**(3), 821–827 (2005)
4. Assa, J., Caspi, Y., Cohen-Or, D.: Action synopsis: Pose selection and illustration. ACM Transactions on Graphics (SIGGRAPH) **24**(3), 667–676 (2005)
5. Axelrod, A.: Msc thesis. Dept. of Computer Science Tel Aviv University (July, 2006)
6. Axelrod, A., Caspi, Y., Gamliel, A., Matsushita, Y.: Interactive video exploration with pose slices. Sketches, SIGGRAPH (2006)
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6), 1373–1396 (2003)
8. BenAbdelkader, C., Cutler, R., Davis, L.: Gait recognition using image self-similarity. EURASIP Journal on Applied Signal Processing **15**(4), 572–585 (2004)
9. Cassinelli, A., Ito, T., Ishikawa, M.: Khronos projector. Emerging Technologies, SIGGRAPH (2005)
10. Chiu, P., Girgensohn, A., Liu, Q.: Stained-glass visualization for highly condensed video summaries. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pp. 2059–2062. IEEE, Taipei, Taiwan (2004)
11. CMU: Motion capture database, cmu graphics lab. http://mocap.cs.cmu.edu/ (2002)
12. CNN: Cable news network. http://www.cnn.com/video/ (2004)
13. Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 53–61. New York, NY, USA (2006)
14. Google: Google video sharing service. http://video.google.com/ (2005)
15. Irani, M., Anandan, P., Hsu, S.: Mosaic based representations of video sequences and their applications. In: International Conference on Computer Vision, pp. 605–611. Washington, DC, USA (1995)
16. Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**, 1–27 (1966)
17. Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. ACM Transactions on Graphics (SIGGRAPH) **24**(3), 595–600 (2005)
18. Liu, F., Zhuang, Y., Wu, F., Pan, Y.: 3D motion retrieval with motion index tree. Computer Vision and Image Understanding **92**(2-3), 265–284 (2003)
19. Loy, G., Sullivan, J., Carlsson, S.: Pose-based clustering in action sequences. In: Workshop on Higher-Level Knowledge in 3D Modeling & Motion Analysis, pp. 66–72. Nice, France (2003)
20. Massey, M., Bender, W.: Salient stills: Process and practice. IBM Systems Journal **35**(3/4), 557–573 (1996)
21. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856. MIT Press, Cambridge, MA (2001)
22. Rav-Acha, A., Pritch, Y., Lischinski, D., Peleg, S.: Dynamosaicing: Video mosaics with non-chronological time. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 58–65. San Diego, CA, USA (2005)
23. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
24. Sun, J., Jia, J., Tang, C.K., Shum, H.Y.: Poisson matting. ACM Transactions on Graphics (SIGGRAPH) **23**(3), 315–321 (2004)
25. Sun, J., Zhang, W., Tang, X., Shum, H.Y.: Background cut. In: European Conference on Computer Vision, pp. 628–641. Graz, Austria (2006)
26. Szeliski, R., Shum, H.Y.: Creating full view panoramic image mosaics and environment maps. ACM Transactions on Graphics (SIGGRAPH) pp. 251–258 (1997)
27. Taniguchi, Y., Akutsu, A., Tonomura, Y.: Panoramaexcerpts: extracting and packing panoramas for video browsing. In: MULTIMEDIA: Proceedings of the fifth ACM international conference on Multimedia, pp. 427–436. ACM Press, New York, NY, USA (1997)
28. Wang, J., Bhat, P., Colburn, R.A., Agrawala, M., Cohen, M.F.: Interactive video cutout. ACM Transactions on Graphics (SIGGRAPH) **24**(3), 585–594 (2005)

Yaron Caspi Yaron Caspi received the BSc degree in mathematics and computer science from the Hebrew University of Jerusalem in 1991 and the MSc degree in mathematics and computer science from the Weizmann Institute of Science in 1993. From 1994-1999, he worked at several computer-vision companies. He received the PhD degree in 2003 from the Weizmann Institute of Science, working on sequence-to-sequence alignment. For his PhD research, he received the Israeli Knesset (Israeli parliament) outstanding student award. He received the best paper award at ECCV 2002, and the honorable mention at ICCV 2001. He is currently at Weizmann Institute of Science.

Anat Axelrod Anat Axelrod is an M.Sc. student at the Computer Science school of Tel-Aviv University and a teaching assistant at the Interdisciplinary Center Herzliya. She received her B.A. degree in Computer Science from the Interdisciplinary Center Herzliya in 2002. Her research interests are computer vision and video visualization. She is also interested in visual arts and had created illustrations for books.

Yasuyuki Matsushita Yasuyuki Matsushita received the BEng, MEng, and PhD degrees in electrical engineering from the University of Tokyo in 1998, 2000, and 2003, respectively. Currently, he is a researcher in Visual Computing Group at Microsoft Research Asia. His research interests include photometric methods in computer vision and video improvement algorithms. He is a member of IEEE.



Alon Gamliel Alon Gamliel received the BSc degree in Computers Engineering from the Technion, Israel Institute of Technology, in 2001. He is currently an MSc student and a teaching assistant at Tel-Aviv University. His research interests are computer graphics, video segmentation and matting techniques.