Making a Long Video Short: Dynamic Video Synopsis *

Alex Rav-Acha Yael Pritch Shmuel Peleg
School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel
E-Mail: {alexis,yaelpri,peleg}@cs.huji.ac.il

Abstract

The power of video over still images is the ability to represent dynamic activities. But video browsing and retrieval are inconvenient due to inherent spatio-temporal redundancies, where some time intervals may have no activity, or have activities that occur in a small image region. Video synopsis aims to provide a compact video representation, while preserving the essential activities of the original video.

We present dynamic video synopsis, where most of the activity in the video is condensed by simultaneously showing several actions, even when they originally occurred at different times. For example, we can create a "stroboscopic movie", where multiple dynamic instances of a moving object are played simultaneously. This is an extension of the still stroboscopic picture.

Previous approaches for video abstraction addressed mostly the temporal redundancy by selecting representative key-frames or time intervals. In dynamic video synopsis the activity is shifted into a significantly shorter period, in which the activity is much denser.

Video examples can be found online in http://www.vision.huji.ac.il/synopsis

1. Introduction

Video synopsis (or abstraction) is a temporally compact representation that aims to enable video browsing and retrieval. We present an approach to video synopsis which optimally reduces the spatio-temporal redundancy in video. As an example, consider the schematic video clip represented as a space-time volume in Fig. 1. The video begins with a person walking on the ground, and after a period of inactivity a bird is flying in the sky. The inactive frames are omitted in most video abstraction methods. Video synopsis is substantially more compact, by playing the person and the bird simultaneously. This makes an optimal use of image regions by shifting events from their original time interval to another time interval when no other activity takes place at this spatial location. Such manipulations relax the chronological consistency of events as was first presented in [14].

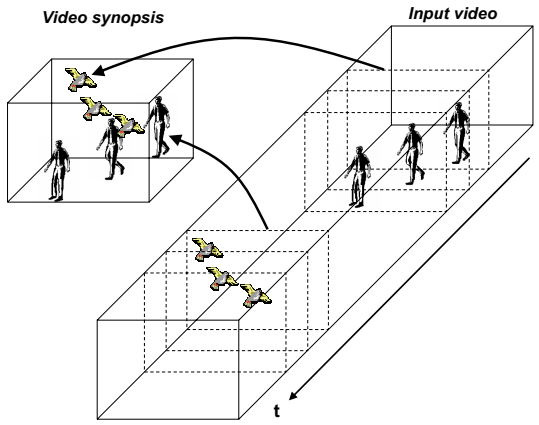


Figure 1. The input video shows a walking person, and after a period of inactivity displays a flying bird. A compact video synopsis can be produced by playing the bird and the person simultaneously.

The dynamic video synopsis suggested in this paper is different from previous video abstraction approaches (reviewed in Sec. 1.1) in the following two properties: (i) The video synopsis is itself a video, expressing the dynamics of the scene. (ii) To reduce as much spatio-temporal redundancy as possible, the relative timing between activities may change. The later point allows for the unique contributions in this paper.

In Sec. 2 we describe a low-level method to produce the synopsis video using optimizations on Markov Random Fields [9].

In Sec. 3 we present an object-based approach in which

*This research was supported by the Israel Science Foundation, Grant number 354/04.

objects are extracted from the input video. Similar moving object detection was also done in other object-based video summary methods [7, 5, 16]. However, these methods use object detection for identifying significant key frames and do not combine activities from different time intervals. The detection of moving objects, as was used in our experiments is described in Sect. 1.2.

One of the options presented in this work is the ability to display multiple dynamic appearances of a single object. This effect is a generalization of the "stroboscopic" pictures used in traditional video synopsis of moving objects [6, 1].

Since this work presents a video-to-video transformation, the reader is encouraged to view the video examples in *http://www.vision.huji.ac.il/synopsis*.

1.1. Related Work on Video Abstraction

There are two main approaches for video synopsis (or video abstraction). In one approach, a set of salient images (key frames) is selected from the original video sequence. The key frames that are selected are the ones that best represent the video [7, 18]. In another approach a collection of short video sequences is selected [15]. The second approach is less compact, but gives a better impression of the scene dynamics. Those approaches (and others) are described in comprehensive surveys on video abstraction [10, 11].

In both approaches above, entire frames are used as the fundamental building blocks. A different methodology uses mosaic images together with some meta-data for video indexing [6, 13, 12]. In this case the static synopsis image includes objects from different times.

1.2. Activity Detection

This work assumes that every input pixel has been labeled with its level of "activity". Evaluation of the activity level is out of the scope of our work, and can be done using one of various methods for detecting irregularities [4, 17], moving object detection, and object tracking.

We have selected for our experiments a simple and commonly used activity indicator, where an input pixel $I(x, y, t)$ is labeled as "active" if its color difference from the temporal median at location (x, y) is larger than a given threshold.

Active pixels are defined by the characteristic function

$$\chi(p) = \begin{cases} 1 & if\ p\ is\ active \\ 0 & otherwise, \end{cases}$$

To clean the activity indicator from noise, a median filter is applied to χ before continuing with the synopsis process.

While it is possible to use a continuous activity measure, we have concentrated in this paper on the binary case. A continuous activity measure can be used with almost all equations in this paper with only minor changes.

2. Video Synopsis by Energy Minimization

Let N frames of an input video sequence be represented in a 3D space-time volume $I(x, y, t)$, where (x, y) are the spatial coordinates of this pixel, and $1 \leq t \leq N$ is the frame number.

We would like to generate a synopsis video $S(x, y, t)$ having the following properties:

- The video synopsis S should be substantially shorter than the original video I.
- Maximum "activity" from the original video should appear in the synopsis video.
- The motion of objects in the video synopsis should be similar to their motion in the original video.
- The video synopsis should look good, and visible seams or fragmented objects should be avoided.

The synopsis video S having the above properties is generated with a mapping M, assigning to every coordinate (x, y, t) in the synopsis S the coordinates of a source pixel from I. We focus in this paper on time shift of pixels, keeping the spatial locations fixed. Thus, any synopsis pixel $S(x, y, t)$ can come from an input pixel $I(x, y, M(x, y, t))$. The time shift M is obtained by solving an energy minimization problem, where the cost function is given by

$$E(M) = E_a(M) + \alpha E_d(M), \tag{1}$$

where $E_a(M)$ indicates the loss in activity, and $E_d(M)$ indicates the discontinuity across seams. The loss of activity will be the number of active pixels in the input video I that do not appear in the synopsis video S,

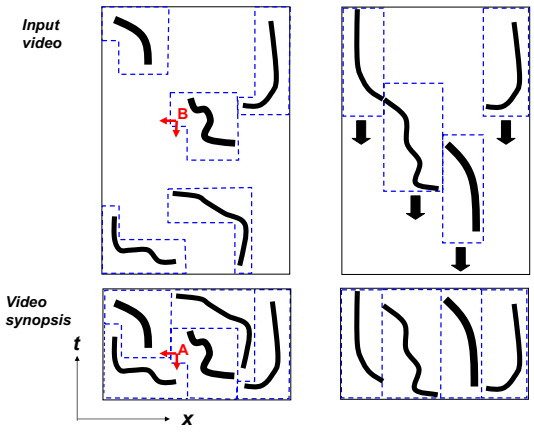
$$E_a(M) = \sum_{(x,y,t)\in I} \chi(x, y, t) - \sum_{(x,y,t)\in S} \chi(x, y, M(x, y, t)). \tag{2}$$

The discontinuity cost E_d is defined as the sum of color differences across seams between spatiotemporal neighbors in the synopsis video and the corresponding neighbors in the input video (A similar formulation can be found in [1]):

$$E_d(M) = \sum_{(x,y,t)\in S} \sum_i \| S(\,(x, y, t) + e_i\,) - I(\,(x, y, M(x, y, t)) + e_i\,) \|^2 \tag{3}$$

where e_i are the six unit vectors representing the six spatio-temporal neighbors. A demonstration of the space-time operations that create a short video synopsis by minimizing the cost function (1) is shown in Fig. 2.a.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE



(a) (b)

Figure 2. In this space-time representation of video, moving objects created the "activity strips". The upper part represents the original video, while the lower part represents the video synopsis. (a) The shorter video synopsis S is generated from the input video I by including most active pixels. To assure smoothness, when pixel A in S corresponds to pixel B in I, their "cross border" neighbors should be similar.
(b) Consecutive pixels in the synopsis video are restricted to come from consecutive input pixels.

Notice that the cost function $E(M)$ (Eq. 1) corresponds to a 3D Markov random field (MRF) where each node corresponds to a pixel in the 3D volume of the output movie, and can be assigned any time value corresponding to an input frame. The weights on the nodes are determined by the activity cost, while the edges between nodes are determined according to the discontinuity cost. The cost function can therefore be minimized by algorithms like iterative graph-cuts [9].

2.1. Restricted Solution Using a 2D Graph

The optimization of Eq. (1), allowing each pixel in the video synopsis to come from any time, is a large-scale problem. For example, an input video of 3 minutes which is summarized into a video synopsis of 5 seconds results in a graph with approximately 2^{25} nodes, each having 5400 labels.

It was shown in [2] that for cases of dynamic textures or objects that move in horizontal path, 3D MRFs can be solved efficiently by reducing the problem into a 1D problem. In this work we address objects that move in a more general way, and therefore we use different constraints. Consecutive pixels in the synopsis video S are restricted to come from consecutive pixels in the input video I. Under this restriction the 3D graph is reduced to a 2D graph where each node corresponds to a spatial location in the synopsis movie. The label of each node $M(x, y)$ determines the frame number t in I shown in the first frame of S, as illustrated in Fig. 2.b. A seam exists between two neighboring locations (x_1, y_1) and (x_2, y_2) in S if $M(x_1, y_1) \neq M(x_2, y_2)$, and the discontinuity cost $E_d(M)$ along the seam is a sum of the color differences at this spatial location over all frames in S.

(a) (b)

Figure 3. The activity in a surveillance video can be condensed into a much shorter video synopsis. (a) A typical frame from the original video. (b) A frame from the video synopsis.

$$E_d(M) = \sum_{x,y} \sum_{i} \sum_{t=1}^{K} || \, S((x, y, t) + e_i) - I((x, y, M(x, y) + t) + e_i) \, ||^2 \quad (4)$$

where e_i are now four unit vectors describing the four spatial neighbors.

The number of labels for each node is $N - K$, where N and K are the number of frames in the input and output videos respectively. The activity loss for each pixel is:

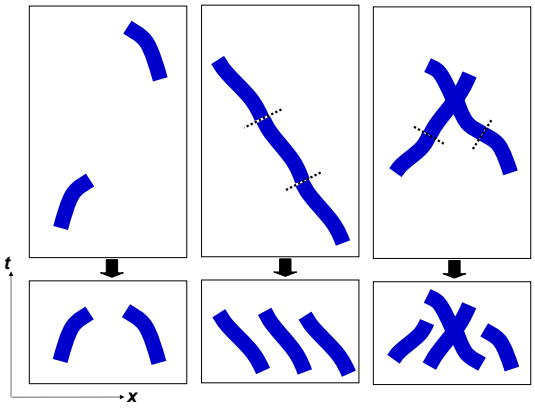
$$E_a(M) = \sum_{x,y} (\sum_{t=1}^{N} \chi(x, y, t) - \sum_{t=1}^{K} \chi(x, y, M(x, y) + t)).$$

Fig. 3 shows an original frame, and a frame from a synopsis video that was obtained using this restricted solver.

3. Object-Based Synopsis

The low-level approach for dynamic video synopsis as described earlier is limited to satisfying local properties such as avoiding visible seams. Higher level object-based properties can be incorporated when objects can be detected. For example, avoiding the stroboscopic effect requires the detection and tracking of each object in the volume. This section describes an implementation of object-based approach for dynamic video synopsis. Several object-based video summary methods exist in the literature (for example [7, 5, 16]), and they all use the detected objects for the selection of significant frames. Unlike these methods, we shift objects in time and create new synopsis frames that never appeared in the input sequence in order to make a better use of space and time.

Moving objects are detected as described in Sec. 1.2 by comparing each pixel to the temporal median and threshold-



(a) (b) (c)

Figure 4. A few examples for a schematic temporal rearrangement of objects. Moving or active objects created the "activity strips". The upper parts represents the original video, and the lower parts represent the video synopsis.
(a) Two objects recorded at different times are shifted to the same time interval in the video synopsis.
(b) A single object moving during a long period is broken into segments having a shorter time intervals, and those are played simultaneously creating a dynamic stroboscopic effect.
(c) Intersection of objects does not disturb the synopsis when object volumes are broken into segments.

ing this difference. This is followed by noise cleaning using a spatial median filter, and by grouping together spatio-temporal connected components. This process results in a set of objects, where each object b is represented by its characteristic function

$$\chi_b(x,y,t) = \begin{cases} 1 & if \ \ (x,y,t) \in b \\ 0 & otherwise, \end{cases} \tag{5}$$

From each object, segments are created by selecting subsets of frames in which the object appears. Such segments can represent different time intervals, optionally taken at different sampling rates.

The video synopsis S will be constructed from the input video I using the following steps:

1. Objects $b_1 \ldots b_r$ are extracted from the input video I.
2. A set of non-overlapping segments B is selected from the original objects.
3. A temporal shift M is applied to each selected segment, creating a shorter video synopsis while avoiding occlusions between objects and enabling seamless stitching. This is explained in Fig. 1 and Fig. 4. An example is shown in Fig. 5.

Steps 2 and 3 above are inter-related, as we would like to select the segments and shift them in time to obtain a short and seamless video synopsis.

In the object based representation, a pixel in the resulting synopsis may have multiple sources (coming from different objects) and therefore we add a post-processing step in which all objects are stitched together. The background image is generated by taking a pixel's median value over all the frames of the sequence. The selected objects can then be blended in, using weights proportional to the distance (in RGB space) between the pixel value in each frame and the median image. This stitching mechanism is similar to the one used in [6].

We define the set of all pixels which are mapped to a single synopsis pixel $(x,y,t) \in S$ as $src(x,y,t)$, and we denote the number of (active) pixels in an object (or a segment) b as $\#b = \sum_{x,y,t \in I} \chi_b(x,y,t)$.

We then define an energy function which measures the cost for a subset selection of segments B and for a temporal shift M. The cost includes an activity loss E_a, a penalty for occlusions between objects E_o and a term E_l penalizing long synopsis videos:

$$E(M,B) = E_a + \alpha E_o + \beta E_l \tag{6}$$

where

$$\begin{aligned} E_a &= \sum_b \#b - \sum_{b \in B} \#b \\ E_o &= \sum_{(x,y,t) \in S} Var\{src(x,y,t)\} \\ E_l &= length(S) \end{aligned} \tag{7}$$

In order to be able to process videos even when the segmentation of moving objects is not perfect, we have penalized occlusions instead of totally preventing them. This occlusion penalty enables flexibility in temporal arrangement of the objects, even when the segmentation is not perfect, and pixels of an object may include some background.

Additional term can be added, which bias the temporal ordering of the synopsis video towards the ordering of the input video.

Minimizing the above energy over all possible segment-selections B and a temporal shift M is very exhaustive due to the large number of possibilities. However, the problem can be scaled down significantly by restricting the solutions. Two restricted schemes are described in the following sections.

3.1. Video-Synopsis with a Pre-determined Length

In this paragraph we describe the case where a short synopsis video of a predetermined length K is constructed from a longer video. In this scheme, each object is partitioned into overlapping and consecutive segments of length K. All the segments are time-shifted to begin at time $t = 1$, and we

Figure 5. One frame from a video synopsis with the dynamic stroboscopic effect as illustrated in Fig. 4.b. The video is in *http://www.vision.huji.ac.il/synopsis*.

are left with deciding which segments to include in the synopsis video. Obviously, with this scheme some objects may not appear in the synopsis video.

We first define an occlusion cost between all pairs of segments. Let b_i and b_j be two segments with appearance times t_i and t_j, and let the support of each segment be represented by its characteristic function χ (as in Eq.5).

The cost between these two segments is defined to be the sum of color differences between the two segments, after being shifted to time $t = 1$.

$$v(b_i, b_j) = \sum_{x,y,t \in S} (I(x, y, t + t_i) - I(x, y, t + t_j))^2 \cdot \quad (8)$$
$$\cdot \chi_{b_i}(x, t, t + t_i) \cdot \chi_{b_j}(x, t, t + t_j)$$

For the synopsis video we select a partial set of segments B which minimizes the cost in Eq. 6 where now E_l is constant K, and the occlusion cost is given by

$$E_o(B) = \sum_{i,j \in B} v(b_i, b_j) \quad (9)$$

To avoid showing the same spatio-temporal pixel twice (which is admissible but wasteful) we set $v(b_i, b_j) = \infty$ for segments b_i and b_j that intersect in the original movie. In addition, if the stroboscopic effect is undesirable, it can be avoided by setting $v(b_i, b_j) = \infty$ for all b_i and b_j that were sampled from the same object.

Simulated Annealing [8] is used to minimize the energy function. Each state describes the subset of segments that are included in the synopsis, and neighboring states are taken to be sets in which a segment is removed, added or replaced with another segment.

(a) (b)

(c)

Figure 6. An example when a short synopsis can describe a longer sequence with no loss of activity and without the stroboscopic effect. Three objects can be time shifted to play simultaneously. (a) The schematic space-time diagram of the original video (top) and the video synopsis (bottom). (b) Three frames from original video. (c) One frame from the synopsis video.

After segment selection, a synopsis movie of length K is constructed by pasting together all the shifted segments. An example of a synopsis using this approach is given in Fig. 5

3.2. Lossless Video Synopsis

For some applications, such as video surveillance, we may prefer a longer synopsis video, but in which all activities are guaranteed to appear. In this case, the objective is not to select a set of object segments as was done in the previous section, but rather to find a compact temporal rearrangement of the object segments.

Again, we use Simulated Annealing to minimize the energy. In this case, a state corresponds to a set of time shifts for all segments, and two states are defined as neighbors if their time shifts differ for only a single segment. There are two issues that should be notes in this case:

- Object segments that appear in the first or last frames should remain so in the synopsis video. (otherwise they may suddenly appear or disappear). We take care that each state will satisfy this constraint by fixing the

Figure 7. When a camera tracks the running lioness, the synopsis video is a panoramic mosaic of the background, and the foreground includes several dynamic copies of the running lioness.

temporal shifts of all these objects accordingly.

- The temporal arrangement of the input video is commonly a local minima of the energy function, and therefore is not a preferable choice for initializing the Annealing process. We initialized our Simulated Annealing with a shorter video, were all objects overlap.

An example of a synopsis using this approach is given in Fig. 6

4. Panoramic Video Synopsis

When a video camera is scanning a scene, much redundancy can be eliminated by using a panoramic mosaic. Yet, existing methods construct a single panoramic image, in which the scene dynamics is lost. Limited dynamics can be represented by a stroboscopic image [6, 1, 3], where moving objects are displayed at several locations along their paths.

A panoramic synopsis video can be created by simultaneously displaying actions that took place at different times in different regions of the scene. A substantial condensation may be obtained, since the duration of activity for each object is limited to the time it is being viewed by the camera. A special case is when the camera tracks an object (such as the running lioness shown in Fig. 7). In this case, a short video synopsis can be obtained only by allowing the Stroboscopic effect.

Constructing the panoramic video synopsis is done in a similar manner to the regular video synopsis, with a preliminary stage of aligning all the frames to some reference frame.

5. Surveillance Examples

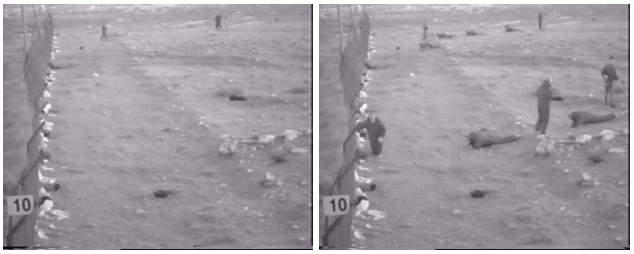
An interesting application for video synopsis may be the access to stored surveillance videos. When it becomes necessary to examine certain events in the video, it can be done much faster with video synopsis.

Fig. 6 gives an example of the power of video synopsis in condensing all activity into a short period, without losing any activity. This was done using a video collected from a camera monitoring a coffee station. Two additional examples are given from real surveillance cameras. Fig. 8 uses a video captured by a camera watching a city street, with

(a)

(b) (c)

Figure 8. Video synopsis from street surveillance. (a) A typical frame from the original video (22 seconds). (b) A frame from a video synopsis movie (2 seconds) showing condensed activity. (c) A frame from a shorter video synopsis (0.7 seconds), showing an even more condensed activity.



(a) (b)

Figure 9. Video synopsis from fence surveillance. (a) A frame from the original video. (b) A frame from a video synopsis, showing simultaneously several occurrences of the crawling and walking soldiers.

pedestrians occasionally crossing the field of view. Many of them can be collected into a very condensed synopsis. In Fig. 9 video synopsis is applied to video from a camera monitoring a security fence. There is very little activity near the fence, and from time to time we can see a soldier crawling towards the fence. The video synopsis shows all instances of crawling and walking soldiers simultaneously, or optionally making the synopsis video even shorter by playing it stroboscopically.

6. Video Indexing Through Video Synopsis

Video synopsis can be used for video indexing, providing the user with efficient and intuitive links for accessing actions in videos. This can be done by associating with every synopsis pixel a pointer to the appearance of the corresponding object in the original video.

In video synopsis, the information of the video is projected into the "space of activities", in which only activities matter, regardless of their temporal context (although we still preserve the spatial context). As activities are concentrated in a short period, specific activities in the video can be accessed with ease.

7. Discussion

Video synopsis has been proposed as an approach for condensing the activity in a video into a very short time period. This condensed representation can enable efficient access to activities in video sequences.

Two approaches were presented: one approach uses low-level graph optimization, where each pixel in the synopsis video is a node in this graph. This approach has the benefit of obtaining the synopsis video directly from the input video, but the complexity of the solution may be very high. An alternative approach is to first detect moving objects, and perform the optimization on the detected objects. While a preliminary step of motion segmentation is needed in the second approach, it is much faster, and object based constraints are possible.

The activity in the resulting video synopsis is much more condensed than the activity in any ordinary video, and viewing such a synopsis may seem awkward to the non experienced viewer. But when the goal is to observe much information in a short time, video synopsis delivers this goal. Special attention should be given to the possibility of obtaining dynamic stroboscopy. While allowing a further reduction in the length of the video synopsis, dynamic stroboscopy may need further adaptation from the user. It does take some training to realize that multiple spatial occurrences of a single object indicate a longer activity time.

While we have detailed a specific implementation for dynamic video synopsis, many extensions are straight forward. For example, rather than having a binary "activity" indicator, the activity indicator can be continuous. A continuous activity can extend the options available for creating the synopsis video, for example by controlling the speed of the displayed objects based on their activity levels.

Video synopsis may also be applied for long movies consisting of many shots. Theoretically, our algorithm will not join together parts from different scenes due to the occlusion (or discontinuity) penalty. In this case the simple background model used for a single shot has to be replaced with an adjustable background estimator. Another approach that can be applied in long movies is to use an existing method for shot boundary detection and create video synopsis on each shot separately.

References

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *SIGGRAPH*, pages 294–302, 2004.

[2] A. Agarwala, K. C. Zheng, C. Pal, M. Agrawala, M. Cohen, B. Curless, D. Salesin, and R. Szeliski. Panoramic video textures. In *SIGGRAPH*, pages 821–827, 2005.

[3] J. Assa, Y. Caspi, and D. Cohen-Or. Action synopsis: Pose selection and illustration. In *SIGGRAPH*, pages 667–676, 2005.

[4] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *ICCV*, pages I: 462–469, Beijing, 2005.

[5] A. M. Ferman and A. M. Tekalp. Multiscale content extraction and representation for video indexing. *Proc. of SPIE*, 3229:23–31, 1997.

[6] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing: Image Communication*, 8(4):327–351, 1996.

[7] C. Kim and J. Hwang. An integrated scheme for object-based video abstraction. In *ACM Multimedia*, pages 303–311, New York, 2000.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 4598(13):671–680, 1983.

[9] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *ECCV*, pages 65–81, 2002.

[10] Y. Li, T. Zhang, and D. Tretter. An overview of video abstraction techniques. Technical Report HPL-2001-191, HP Laboratory, 2001.

[11] J. Oh, Q. Wen, J. lee, and S. Hwang. Video abstraction. In S. Deb, editor, *Video Data Mangement and Information Retrieval*, pages 321–346. Idea Group Inc. and IRM Press, 2004.

[12] C. Pal and N. Jojic. Interactive montages of sprites for indexing and summarizing security video. In *Video Proceedings of CVPR05*, page II: 1192, 2005.

[13] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video abstraction: Summarizing video content for retrieval and visualization. In *Signals, Systems and Computers*, pages 915–919, 1998.

[14] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg. Dynamosaics: Video mosaics with non-chronological time. In *CVPR*, pages 58–65, Washington, DC, 2005.

[15] A. M. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. In *CAIVD*, pages 61–70, 1998.

[16] A. Stefanidis, P. Partsinevelos, P. Agouris, and P. Doucette. Summarizing video datasets in the spatiotemporal domain. In *DEXA Workshop*, pages 906–912, 2000.

[17] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, pages 819–826, 2004.

[18] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Syst.*, 10(2):98–115, 2004.