CRAM: Compact Representation of Actions in Movies

Mikel Rodriguez
Computer Vision Lab, University of Central Florida. Orlando, FL
mikel@cs.ucf.edu

Abstract

Thousands of hours of video are recorded every second across the world. Due to the fact that searching for a particular event of interest within hours of video is time consuming, most captured videos are never examined, and are only used in a post-factum manner. In this work, we introduce activity-specific video summaries, which provide an effective means of browsing and indexing video based on a set of events of interest. Our method automatically generates a compact video representation of a long sequence, which features only activities of interest while preserving the general dynamics of the original video. Given a long input video sequence, we compute optical flow and represent the corresponding vector field in the Clifford Fourier domain. Dynamic regions within the flow field are identified within the phase spectrum volume of the flow field. We then compute the likelihood that certain activities of relevance occur within the the video by correlating it with spatio-temporal maximum average correlation height filters. Finally, the input sequence is condensed via a temporal shift optimization, resulting in a short video clip which simultaneously displays multiple instances of each relevant activity.

1. Introduction

Every day millions of hours of video are captured around the world by CCTV cameras, webcams, and traffic-cams. In the United States alone, an estimated 26 million video cameras spit out more than four billion hours of video footage every week. In the time it takes to read this sentence, close to 20,000 hours of video have been captured and saved at different locations in the U.S. However, the vast majority of this wealth of data is never analyzed by humans. Instead, most of the video is used in an archival, post-factum manner once an event of interest has occurred.

The main reason for this lack of exploitation resides in the fact that video browsing and retrieval are inconvenient due to inherent spatio-temporal redundancies, in which extended periods of time contain little to no activities or events of interest. In most videos, a specific activity of interest may only occur in a relatively small region along the entire spatio-temporal extent of the video.

There exists a large body of work that addresses the topic of activity recognition which focuses mainly on detection in short pre-segmented video clips commonly found in publicly available, standard action datasets. In this work, we attempt to move beyond only performing action detection in an effort to provide a means of generating a compact video representation based on a set of activities of interest, while preserving the scene dynamics of the original video. In our approach, a user specifies which activities interest him and the video is automatically condensed to a short clip which captures the most relevant events based on the user's preference. We follow the output summary video format of non-chronological video synopsis approaches, in which different events which occur at different times may be displayed concurrently, even though they never occur simultaneously in the original video. However, instead of assuming that all moving objects are interesting, priority is given to specific activities of interest which pertain to a user's query. This provides an efficient means of browsing through large collections of video for events of interest.

2. Related Work

Action recognition and event classification in video have been studied extensively in recent years; a comprehensive review can be found in surveys on the topic [9, 1]. Most of the existing work can be broadly categorized into approaches based on tracking [18, 4], interest points [11], geometrical models of human body parts [6], 3D information [13], volumetric space-time shapes [21], action clustering [10] and temporal-templates [2].

A common theme in all of these approaches is their focus on detection. That is, given a learned model of an action class, emphasis is placed on detecting instances of the learned action within small testing clips typically found in standard action datasets. After performing detection, most methods do not go beyond placing a bounding box delimiting the spatio-temporal extent of the detected action. Our present work aims at moving beyond detection by examining the role of action recognition in efficient video repre-

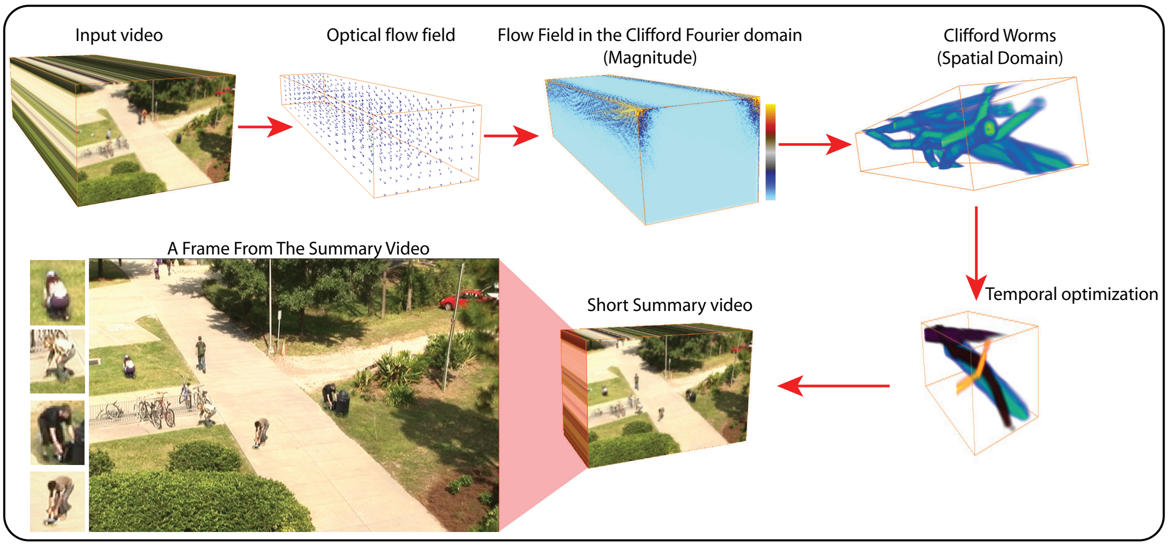
Input video
Optical flow field
Flow Field in the Clifford Fourier domain (Magnitude)
Clifford Worms (Spatial Domain)
Temporal optimization
Short Summary video
A Frame From The Summary Video

Figure 1. A frame from a video summary for the "picking up" action, along with the various steps of the action-specific video synopsis process. Given a long input video sequence (spatio-temporal volume), we compute optical flow and represent the corresponding flow field in the Clifford Fourier domain. Dynamic regions (Clifford worms) are identified within the Clifford domain, and a temporal optimization shifts worms which contain activities of interest in the temporal domain to obtain a compact representation of the original video. Finally, we see the resulting short clip which contains four instances of the "picking up" action of interest.

sentations. More specifically, we are interested in the generation of compact video summaries which contain specific actions of interest.

Existing work on compact video representation and summarization can be roughly categorized into static representations (which include key-frame-based methods [7, 22] and mosaic-based methods [15]), and full motion video representations (which includes non-chronological video synopsis [16], bi-directional similarity [20] and smart fast-forward approaches [14]).

Most of these approaches are geared towards providing a compact representation of a video as a whole and do not distinguish between different classes of events. Therefore most of the existing approaches are best suited for uncrowded scenes that contain periods of inactivity and where events are sparse. We are interested in a compact representation of long videos which is based on specific actions of interest. Therefore, it may not be appropriate to rely on static, frame-based or mosaic-based representations of video, given that important events and actions which can only be distinguished upon inspecting a sequence of frames are lost in these static representations. Furthermore, generating a compact video representation based on all moving objects in the scene may lead to the inclusion of irrelevant moving objects. This is especially true in crowded scenes were moving objects abound. Therefore, in this work, we explore the role of action recognition as a means of generating condensed representations of long videos which can efficiently convey only important events and actions of interest which occur over a long period of time.

3. Compact Action-based Video Representation

Our approach to generating compact action-specific video representations is composed of three main phases. First, we begin by determining a set of regions in space-time which contain dynamic objects. Subsequently, we detect specific activities and actions of interest within the long video sequence. Finally, we select dynamic regions which contain events of interest and optimize the temporal shifts of the video summary via an energy minimization. In the following sections we describe each of these steps in more detail.

3.1. Motion Representation

In this section we describe how we identify dynamic regions of a video sequence as potential candidate spatio-temporal locations to be included in the final video summary. For this purpose we begin by computing optical flow for the entire sequence using the flow estimation method described in [12]. However, instead of identifying dynamic regions within the optical flow in the spatial domain, we efficiently identify such regions in the frequency domain.

Given the fact that we seek to identify salient regions within a 3D optical flow field, where at each point we have more than one component (d_x, d_y, magnitude), we cannot employ the traditional Fourier transform which is defined on scalar values as a valid representation without losing any information. In order to efficiently analyze a video sequence in the frequency domain we require an analog to

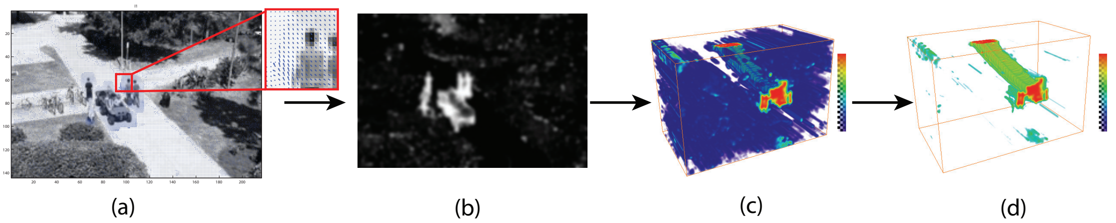

Figure 2. (a) The optical flow field for a long video sequence. (b) A 2D slice of the phase spectrum volume (PSV). (c) A 3D segment of the PSV, high values indicate dynamic regions within the flow field. (d) Candidate dynamic regions (worms)

the classical Fourier transform for vector fields. For this purpose, we follow the framework proposed in [3], in that we apply an algebraic extension to the degrees of freedom of a multi-dimensional Fourier transform by embedding the spectral domain into a domain of Clifford numbers. This class of Fourier transform is commonly referred to as the "Clifford Fourier transform." Using this embedding, we preserve the full information of relative directions of our vector field while identifying potential regions in space-time which should be included in the summary video.

The Clifford Fourier transform (CFT) for multivectors-valued functions in 3D is defined as:

$$\mathcal{F}\{\mathbf{F}\}(\mathbf{u}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|, \qquad (1)$$

where $\mathbf{i}_3$ represents the analog of a complex number in Clifford algebra, such that $\mathbf{i}_3 = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$ and ${\mathbf{i}_3}^2 = -1$. The inverse transform is given by:

$$\mathcal{F}^{-1}\{\mathbf{F}\}(\mathbf{x}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|. \qquad (2)$$

3.2. Dynamic Spatio-temporal Regions

Given a long video sequence, we compute optical flow vectors and magnitude resulting in a 3D vector field. We employ the Clifford embedding described in section 3.1 by performing a 3D Clifford Fourier transform. In order to identify regions of potential activity of interest and shift them in time for an action-specific video summary, we seek to to carve out a set of spatio-temporal regions, or "worms," from the input flow field which suggest areas of dynamic events. Each worm is, in fact, an object, or group of objects, which carves out a spatio-temporal volume as it moves across the scene over time.

It is well known that the amplitude spectrum specifies how much of each sinusoidal component is present in a signal [5] and the phase information specifies where each of the sinusoidal components resides within the signal, which in our domain corresponds to a flow field. Locations within the flow field which have less periodicity or less homogeneity represent potential dynamic regions of interest in the reconstruction of the flow field, which indicates the location of the worm candidates.

Knowing that the phase spectrum of a flow field in the frequency domain can provide insight as to where dynamic events are occurring in the original video, we identify a set of candidate regions in space-time as follows:

Given a flow field $(\mathbf{u})$ of an input video:

$$\begin{aligned} f(x,y,t) &= \mathcal{F}\{\mathbf{F}\}(\mathbf{u}) && (3) \\ p(x,y,t) &= P(f(x,y,t)) && (4) \\ W(x,y,t) &= g(x,y,t) * \left|\left|\mathcal{F}^{-1}\{\mathbf{F}\}\left[e^{\mathbf{i}_3 p(x,y,t)}\right]\right|\right| && (5) \end{aligned}$$

where $\mathcal{F}\{\mathbf{F}\}$ and $\mathcal{F}^{-1}\{\mathbf{F}\}$ denote the Clifford Fourier Transform and inverse Clifford Fourier Transform respectively. $P(f(x,y,t))$ represents the phase spectrum of the vector field which is composed of the u and v components of optical flow as well as the magnitude. $g(x,y,t)$ is a 3D gaussian filter (we use $\sigma = 6$). We obtain potential dynamic regions in the video by converting the phase spectrum into the spatial domain and convolving the resulting scalar field with a gaussian.

The phase spectrum volume (Figure 2-c) contains the "innovation" of a specific region in the flow field. Using the Inverse Clifford Fourier Transform, we can construct the output volume which contains primarily the non-trivial, or unexpected spatio-temporal regions of the flow field, where we expect to find events of interest.

Given the phase spectrum volume we threshold the saliency values and segment out a set of "worms" belonging to different objects which trace some movement across the scene over time. We use the normalized cuts toolbox of the algorithm described in [19] to obtain the tightest clusters in this space-time volume. Each spatio-temporal location in the phase spectrum volume (in the Spatial Domain) which is above a threshold forms a node of a completely connected graph. Edge weights are assigned using the Euclidean distance between connected nodes. Using normalized cuts on this graph we obtain the optimum clustering of dynamic regions in the flow field into a set of worms which can then be shifted in time to generate a summary video. In the next section we describe how we narrow the pool of potential worms to be included in the final summary video by detecting specific activities and actions of interest.

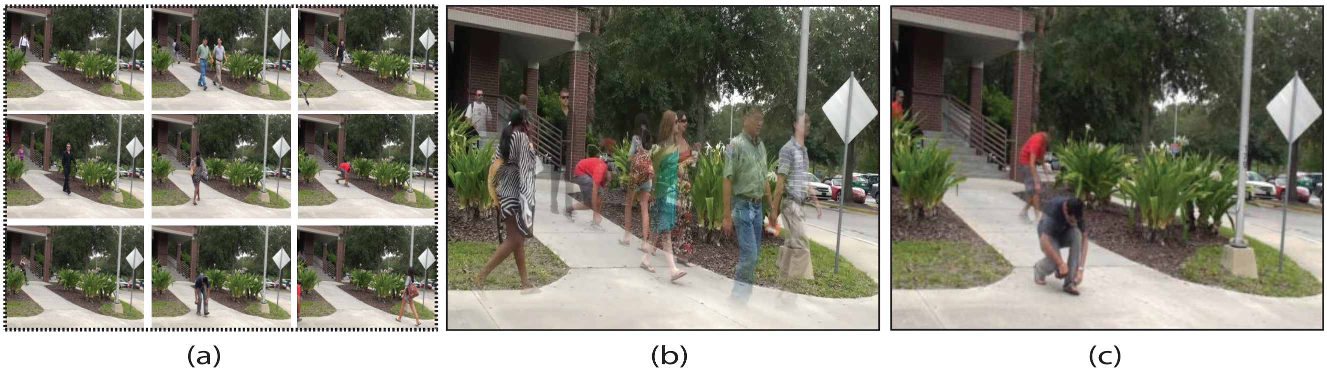

Figure 3. (a) Frames from the original long video sequence. (b) A non-action-based video summary. (c) An action specific video summary based on the "pickup" action of interest.

3.3. Action-Specific Summary

In this work we are interested in compact action-specific video representations. For example, in a parking lot scene, we may be interested in a brief video clip containing all the people entering cars during some time period. Similarly, we may be interested in a quick summary video containing all people who were running through a given scene over the course of a week's worth of video.

In order to generate action-specific summaries we identify the most relevant activities based on pre-defined action templates. Worms containing activities of interest are assembled into a synopsis video of a specified temporal length. In order to include as many activities as possible in the short video synopsis, different action instances may be displayed concurrently, even if they originally occurred at different times. The resulting synopsis video will contain events of interest in a short clip which can serve as an index into the original long video by keeping a pointer to the original spatio-temporal location of the event.

3.3.1 Identifying Activities of Interest

We identify dynamic spatio-temporal regions which contain specific activities of interest by correlating worms with a pool of action templates. Actions such as "run," "walk," "open car door," and "load/unload car trunk" are captured using templates synthesized using a recently proposed [17] generalization of the traditional maximum average correlation height filter to video (3D spatiotemporal volume), and vector value data such as optical flow.

Action templates for activities of interest are generated by computing temporal derivatives of a set of examples. Subsequently, these examples are represented in the frequency domain via a 3D Fourier transform. Given the resulting volumes in the Fourier domain, we proceed to convert the resulting 3D matrix into a column vector by concatenating all the columns of the 3D matrix, resulting in a single column-vector (x_i). This process is repeated for each example of an activity of interest. Finally, the template for a given activity of interest can be generated in the Fourier domain by minimizing:

$$h = (\alpha C + \beta D_x + \gamma S_x)^{-1} m_x, \tag{6}$$

where m_x is the mean of all the x_i vectors, and h is the template in vector form in the frequency domain. C is the diagonal noise covariance matrix of size $d \times d$, where d is the total number of elements in x_i vector. Given that we do not have a specific noise model for a scene, we set $C = \sigma^2 I$, where σ is the standard deviation parameter and I is a $d \times d$ identity matrix. D_x is also a $d \times d$ diagonal matrix representing the average power spectral density of the training videos.

Having obtained a one-dimensional template (h) for an activity of interest, we proceed to assemble a complete 3D filter by reshaping and then applying the inverse Fourier transform. The resulting matrix constitutes the template, H, for the particular activity of interest.

Once a set of action templates has been generated, we can proceed to identify regions in the video which contain activities of interest and should therefore be included in a summary video.

We determine the likelihood that a spatio-temporal region contains a specific activity of interest by correlating the corresponding template with input sequence:

$$c(l,m,n) = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} s(l+x, m+y, n+t) H(x,y,t), \tag{7}$$

where s is the spatio-temporal volume of the long input video, H is the action template (h is its Fourier transform). P, Q, and R are the dimensions of the of the spatio-temporal volumes.

As a result of this operation, we obtain a response, c, of size $(P-L+1) \times (Q-M+1) \times (R-N+1)$. We denote this location by (l^*, m^*, n^*). Due to varying illumination

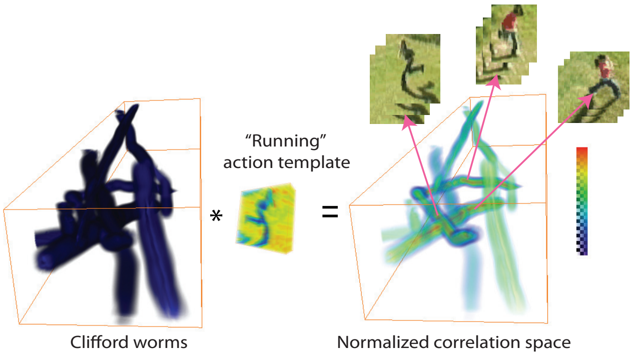

Figure 4. We narrow the pool of potential worms to be included in the final summary video by determining the likelihood that worms contain specific activities and actions of interest.

conditions and noise in the scene, we optimize the response of the filter by normalizing our correlation space:

$$c'(l,m,n) = \frac{c(l,m,n)}{\sqrt{E_H E_S(l,m,n)}}, \quad (8)$$

where $c(l,m,n)$ is given by equation 5. E_H is a scalar value which represents the energy of the filter, and $E_s(l,m,n)$ corresponds to the energy of the test volume at location (l,m,n), given by:

$$E_H = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} H^2(x,y,t), \quad (9)$$

$$E_S(l,m,n) = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} s^2(l+x, m+y, n+t). \quad (10)$$

Each element in the response of the normalized correlation lies within 0 and 1, a fact that can be used as a level of confidence in a pseudo-probabilistic manner to determine which spatio-temporal regions contain events of interest. The peak value in the response of the filter is compared with a threshold (τ). Thresholds for action classes are computed during training as $\tau = \xi * min(p_1, p_2, p_3, ..., p_{N_e})$, where p_i is the peak value obtained from the correlation response when ith training volume was correlated with the 3D MACH filter, ξ is a constant parameter, and N_e is the number of all the training volumes.

3.4. Temporal Shift Optimization

The final synopsis video is generated based on a collection of temporal shifts (S) which map the worms that contain actions of interest to a different time in a summary video such that a more compact representation of the original sequence can be obtained. Given the locations of the action detections, we are able to identify a pool of worms (W) that are likely to contain events of interest. We define an optimal action-based video summary as the one that minimizes the following energy function:

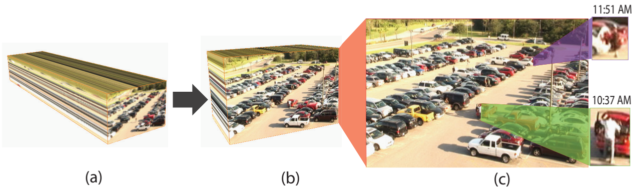

Figure 6. A video summary of "opening trunk" events in a parking lot. (a) A two hour long video sequence is summarized in a one minute clip (b) containing most of the instances of the event of interest ("opening trunk"). The video summary displays multiple instances of the event of interest (which may have occurred at different times) concurrently (c).

$$E(S) = \sum_{w \in W} \alpha E_t(w) + \sum_{w,w' \in W} E_o(w,w') \quad (11)$$

where E_t is the cost associated with the temporal extent of a time shift configuration (maximum temporal location), and E_o is the spatio-temporal overlap cost. The spatio-temporal overlap cost E_o penalizes regions in the video which contain events of interest that are mapped to new temporal locations which results in some degree of overlap between them. It is given by the volume of their space-time overlap. This energy function can be minimized using various optimization techniques, in our experiments we employed simulated annealing as well as the more efficient multi-label graph cut method described in [8]. In the later, labels correspond to time shifts of worms and a cut in the graph represents a specific time shift. The result is an optimal set of time shifts which minimizes the temporal extend of the summary while also minimizing the amount of spatio-temporal overlap between worms.

4. Experiments and Results

We performed a number tests to better understand the ability of the proposed method to cope with a range of video sources. Details about the video sources used in generating the action-specific summaries and the experiments performed are given below. Video samples of our results can also be found in our supplemental material.

4.1. Ground Camera Videos

In the first round of experiments a collection of videos obtained from ground cameras which included parking lot scenes and street scenes was used to generate video summaries of activities of interest. The video corpus contained several hours of video divided across six different clips. Activities of interest in these experiments were defined to be "running," "picking up an object," "entering vehicle," and "loading/unloading trunk."

Each activity of interest occurs multiple times at different points within the collection of long video sequences.

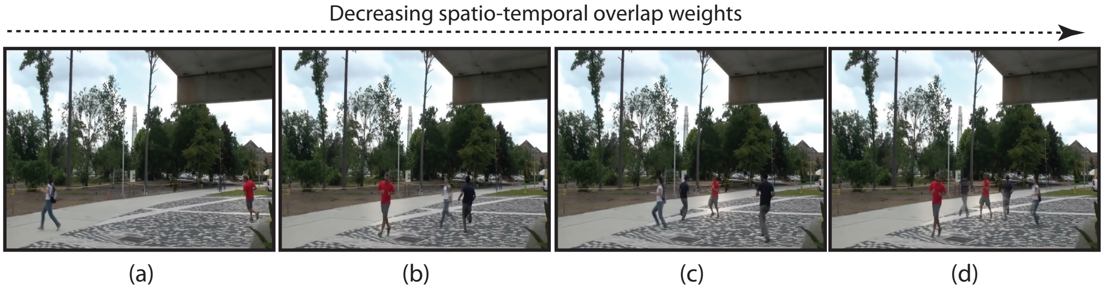

Figure 5. Decreasing the weight of the spatio-temporal overlap cost leads to increasingly compact summaries at the cost of additional overlaps. (a) $\alpha = 0.6$, (b) $\alpha = 0.5$, (c) $\alpha = 0.4$, (d) $\alpha = 0.3$

"Running" occurs 28 times, "picking up an object" occurs 19 times, "entering vehicle" occurs 28 times, and "loading/unloadding trunk" occurs 23 times.

Figure 1 demonstrates the effect of generating a video summary based on the "picking up object" activity of interest. A long video which contains only five instances of the action of interest is represented by a short, one minute clip, containing most of the instances of the event of interest. In this example of our results we see how four different instances of the "picking up object" action are displayed concurrently, despite the fact that they have occurred over a long period of time. The video of this summary and of other results can be found in our supplemental material.

The value of an action-specific video summary is evident in Figure 3. In this experiment we generated a 10 second video summary based on the most dynamic spatio-temporal regions (worms) which results in a short yet cluttered video clip (Figure 3-b) given that all of the spatio-temporal dynamic regions are treated equally. When we employ the action-specific video summary framework using the "picking up" action of interest the resulting clip consists of relevant events (two people picking up an object) and is considerably less cluttered. In long videos of crowded scenes where moving objects abound action-specific video summaries provide a means of distilling the long sequence into a short clip that clearly depicts events of interest that occurred over a period of time. Both of these videos can be found in our supplemental material.

A more challenging scenario is seen in Figure 6, where we have a busy parking lot scene which contains many motions which can potentially be irrelevant to a given user. Therefore, it may not be appropriate to generate a synopsis based on all moving objects in the scene. In this experiment, we generated a video synopsis of a long video clip based on the "open vehicle trunk" event of interest. Despite the fact that instances of the event of interest are relatively small as compared to the rest of the scene, our summary includes seven out of the total eight instances of the event of interest in a one one minute clip. Searching for this particular event manually would require careful observation as the video is fast-forwarded, a time consuming and inefficient process.

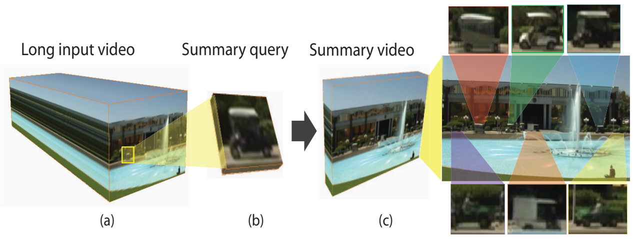

Figure 8. Summary by example: given a long video sequence (a), we specify a spatio-temporal region as a query (b) which contains an event of interest. A video summary (c) which includes events in the scene which match the query is then automatically generated.

A similar cluttered scenario we consider is depicted in Figure 10, where we condense all of the instances of the running action which occur throughout a long traffic sequence into a one minute clip. Given that in this particular scene, running pedestrians tend to occur in one particular region in the video (the crosswalk), we increase the weight of the spatio-temporal overlap cost term (by setting α to 0.3) in order to avoid artifacts caused by multiple running activities which are mapped to the same spatio-temporal region.

The effect of varying the spatio-temporal overlap cost is depicted in our experiment in Figure 5, in which a ten second summary of the "running" action is obtained from a long input video. As we lower the weight (α) of the spatio-temporal overlap cost we observe how additional instances of the running action are included in the video summary resulting in additional clutter.

4.2. Aerial Videos

A second round of experiments was based on aerial video sequences obtained using a UAV equipped with an HD camera mounted on a gimbal. Videos were recorded at a flying altitude of over 400 feet. The collection contains a diverse pool of events such as people getting into vehicles, and people running, which occur over the course of one hour. These videos are divided into sequences which typically average 5-12 minutes in length. In these experiments our goal is to evaluate the ability to generate video synopses based on moving aerial camera video sequences. Given an aerial video sequence obtained by a UAV hovering over a re-

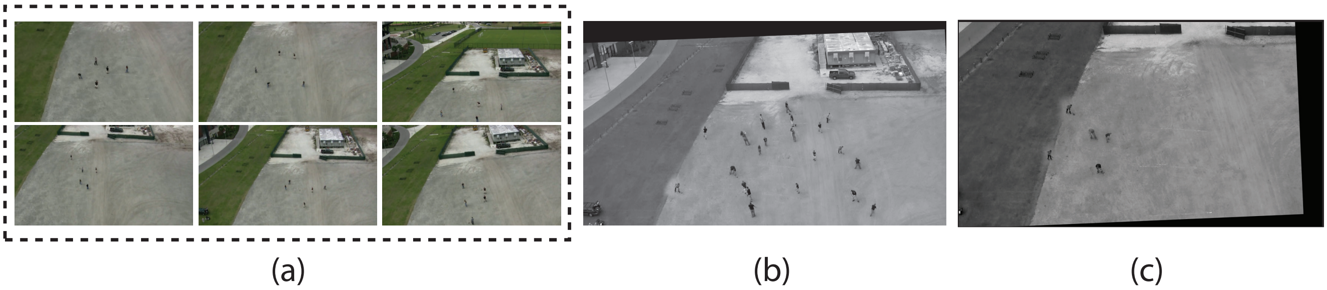

Figure 7. (a) Frames of a long aerial video sequence shot from an R/C helicopter flying at 400 feet. (b) A non-action based video summary.(c) A video summary of the "digging" action.

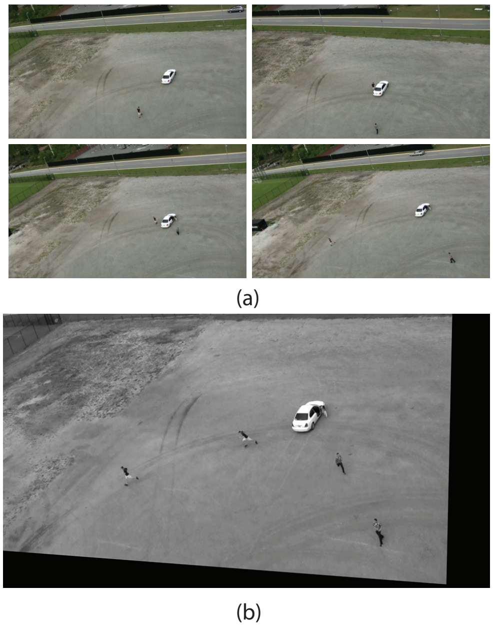

Figure 9. (a) Frames of a 13 minute aerial video sequence. (b) A video summary of the "running" action.

gion of interest, we begin the summary process by performing frame-to-frame registration across the sequence. Subsequently, we identify dynamic regions within the registered aerial video, identify events of interest, and perform temporal optimization using the methods described in Sections 3.3.1, and 3.4 respectively.

Figure 11 depicts an example of a 30 second video summary generated from an aerial video sequence (which is 9 minutes long) based on the "running" action of interest. The short summary clip contains four out of the seven running events which occur in the scene over the entire duration of the longer clip.

Another aerial action video summary is depicted in Figure 7. In this experiment two video summaries are generated from a aerial video. The first is a 30 second non-action video summary. As can be seen in Figure 7-b, this results in a cluttered video clip that contains various movers that originate from different spatio-temporal regions. Figure 7-c depicts an action-based video summary of the same length, that contains five instances of the digging action which occur at different point in time in the original video. Due to small out-of-plane parallax errors which are propagated over time, a modest amount of drift in alignment is accumulated which results in some visible artifacts around some of the shifted action instances.

In our experiments we observed that the main issues related to generating video summaries of aerial sequences are noise in the flow field which is caused by slight errors in the motion compensation. This leads to noisy dynamic regions which are sometimes included in the final summary video. This effect can be observed in Figure 9 in which a ten second video summary of the running action is generated from a 13 minute aerial video. Due to nosy worms individual instances of an action have been segmented into disjoint events which are then shifted in time independently. As can be seen in Figure 9-b two seperate running instances have been segmented into four small running segments which are depicted concurrently in the short summary clip.

4.3. Summary by Example

It is not always possible to obtain a large training set for a collection of events of interest. Nor is it feasible to assume that we are only interested in video summaries of a static set of pre-defined events (such as running, opening car door, etc). Therefore, in our last round of experiments we introduce the concept of "summary by example." That is, given a long video sequence, a user can select any instance of a particular event of interest by specifying a spatial region in the video and the temporal extent of the event. Subsequently, a short video summary which contains all the events that match the selected query is generated for the rest of the long sequence.

Summary by example can be accomplished without any major changes to the overall approach described above. This is due to the fact that we treat the example of the event of interest as a single instance spatio-temporal template. In order to account for the possibility of observing the event of interest at different scales across the long video, we synthesize templates at three scales by resizing the original example. Aside from employing this special case of the spatio-

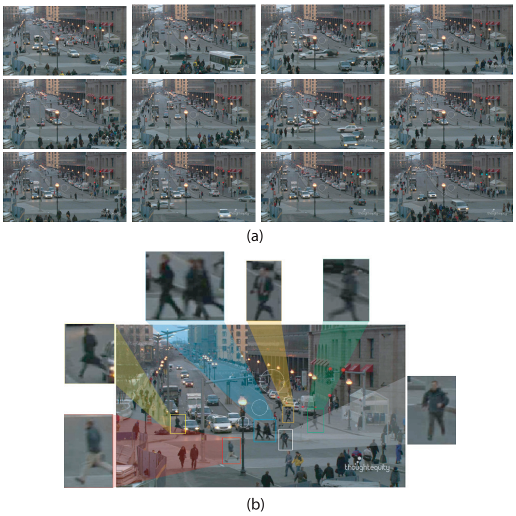

Figure 10. (a)Frames from a long cityscape video. (b) A frame from a short clip generated by our system which captures instances of running in the scene over an extended period of time.

temporal template, the remaining steps of our approach remain the same. Figure 8 depicts a summary by example, in which an event of interest consisting of a moving golf cart is selected within a long video sequence. Based on this example of an event of interest, our method generates a short thirty second video summary which condenses six separate instances of a moving golf cart event which occurs at different times within the long video.

5. Conclusion

We have explored the role of template-based action recognition methods in generating short video summaries of long ground camera videos and aerial videos. We do not consider all moving objects in the long video sequence to be of equal importance when generating a given video summary. Instead we focussed on generating video summaries based on a set of events of interest which can be specified when generating a summary. We found that these activity-specific video summaries provide us with a more meaningful way of quickly reviewing a long video for particular events of interest in the form of a short video clip which condenses all activities of interest that have occurred across some time span. Furthermore, by focusing on events of interest instead of moving objects we were able to generate meaningful summaries of crowded scenes. As future work we intend to explore multi-agent events with long-range spatio-temporal dependencies. We also intend to use confidence values of action detection to draw attention to specific areas in the summary.

5.1. Acknowledgement

This research was funded by Harris Corporation.

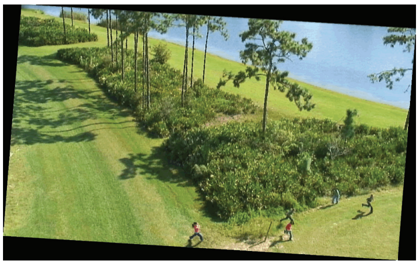

Figure 11. UAV aerial video summary containing the “running” action. Four instances of the running action which occur at different time instances across a long video are displayed concurrently.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. *CVIU*, 73(3), 1999.

[2] A. Bobick and J. Davis. The Recognition of Human Movement Using Temporal Templates. *TPAMI*, 2001.

[3] J. Ebling and G. Scheuermann. Clifford Fourier transform on vector fields. *VCG*, 11(4), 2005.

[4] P. Hong, M. Turk, and T. Huang. Gesture modeling and recognition using finite state machines. In *ICGR*, 2000.

[5] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[6] H. Jiang, M. Drew, and Z. Li. Successive Convex Matching for Action Detection. In *CVPR*, 2006.

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

[8] V. Kolmogorov and R. Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *TPAMI*, 2004.

[9] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, 81(3), 2001.

[10] M. Moslemi Naeini. Clustering and visualizing actions of humans and animals using motion features. *MS Thesis*, 2007.

[11] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007.

[12] N. Papenberg, A. Bruhn, and Brox. Highly accurate optic flow computation with theoretically justified warping. *IJCV*, 67(2):141–158, 2006.

[13] V. Parameswaran and R. Chellappa. View invariants for human action recognition. In *CVPR*, volume 2, 2003.

[14] N. Petrovic, N. Jojic, and T. Huang. Adaptive Video Fast Forward. *Multimedia Tools and Applications*, 26(3), 2005.

[15] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video Abstraction. In *ACSS*, 1998.

[16] Y. Pritch, A. Rav-Acha, and S. Peleg. Non-Chronological Video Synopsis and Indexing. *TPAMI*, 2008.

[17] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH. In *CVPR*, 2008.

[18] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the Space of a Human Action. In *ICCV*, 2005.

[19] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *TPAMI*, 2000.

[20] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *CVPR*, 2008.

[21] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *CVPR*, volume 1, 2005.

[22] X. Zhu, X. Wu, J. Fan, A. Elmagarmid, and W. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Systems*, 10(2), 2004.