Video Précis: Highlighting Diverse Aspects of Videos

Nitesh Shroff, *Student Member, IEEE*, Pavan Turaga, *Member, IEEE*, and Rama Chellappa, *Fellow, IEEE*

***Abstract*—Summarizing long unconstrained videos is gaining importance in surveillance, web-based video browsing, and video-archival applications. Summarizing a video requires one to identify key aspects that contain the essence of the video. In this paper, we propose an approach that optimizes two criteria that a video summary should embody. The first criterion, "coverage," requires that the summary be able to represent the original video well. The second criterion, "diversity," requires that the elements of the summary be as distinct from each other as possible. Given a user-specified summary length, we propose a cost function to measure the quality of a summary. The problem of generating a précis is then reduced to a combinatorial optimization problem of minimizing the proposed cost function. We propose an efficient method to solve the optimization problem. We demonstrate through experiments (on KTH data, unconstrained skating video, a surveillance video, and a YouTube home video) that optimizing the proposed criterion results in meaningful video summaries over a wide range of scenarios. Summaries thus generated are then evaluated using both quantitative measures and user studies.**

***Index Terms*—Exemplar selection, K-means, Ncut, shot segmentation, video summarization.**

I. Introduction

RECENT years have witnessed a tremendous increase in multimedia content driven by inexpensive video cameras and the growth of the Internet. The amount of visual data being recorded and accessed has been increasing with the rising popularity of several social networking and video-sharing websites. In several applications, there is a need to gain a quick overview of the contents of a video. As an example, imagine having to browse through an hour of skating video to view all of the diverse clips corresponding to camel spin, axel-jump, etc., by simple linear browsing. Instead, if one could extract/filter a few segments of the original video, such that they are "mutually exclusive and exhaustive," this will result in significant improvement in user experience. Similarly, in video-based security and surveillance systems, analysts spend long hours sifting through large video recordings of parking lots and airports to locate events of interest. The ability to gain a quick overview of diverse aspects of hour-long videos is vital in these applications.

Manuscript received September 03, 2009; revised February 16, 2010 and June 18, 2010; accepted June 23, 2010. Date of publication July 15, 2010; date of current version November 17, 2010. This work was supported in part by the Office of Naval Research under Grant N00014-09-1-0044. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophe De Vleeschouwer.

The authors are with the Department of Electrical and Computer Engineering and the Center for Automation Research in the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: nshroff@umiacs.umd.edu; pturaga@umiacs.umd.edu; rama@umiacs.umd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2058795

Recently, with the advent of improved multimedia technologies (e.g., inexpensive cameras and video hosting websites), unconstrained videos have become commonplace. Consider a few examples of videos shown in Fig. 1. Each of these videos has more than 5000 frames (each video is approximately 4 min in duration). As can be seen, there exists a great deal of diverse information in each of these videos. This calls attention to the problem of optimally selecting "exemplars" from the video so as to obtain a quick overview of the video without losing any detail. Here, we define exemplars as "informative/key segments" of the video selected to provide condensed and succinct representations of the content of a video. An exemplar could be a single frame or a video segment (continuous sequence of frames) of fixed length depending upon the end application.

This ability to produce an abstract of the video, without losing the details, requires addressing two conflicting requirements: 1) the summary should be representative of the video ("exhaustive") and 2) the summary should highlight diverse aspects of the video ("mutually exclusive"). The first criterion suggests that the summary should "cover" most of the video in terms of global representation. The second criterion suggests that the elements of the summary should be as distinct as possible. In this paper, we propose a novel formulation to quantify this tradeoff and subsequently propose an approach for generating a summary that takes into account these two important considerations. We provide a mathematical formulation of these criteria using a unified cost function and propose an algorithm to solve it.

Related Work: The problem of generating the summary given a video has attracted significant attention, especially over the past few years. Several video abstraction systems have been proposed, and good recent surveys with systematic classification of various approaches can be found in [1]–[6]. An overview of the existing approaches is given in Fig. 2. Most existing approaches have relied predominantly on computation and processing of "shots." A shot is a sequence of frames within a continuous capture period, and the transition between two consecutive shots is termed as the shot boundary. These shot boundaries are then used to detect the temporal span of each shot. Thereafter, exemplars within shots are selected using simple techniques such as the first/middle/last frames or merely random sampling. These approaches were motivated from and tuned to genres such as movies and news videos, where the use of shots is an integral part of the video capturing process. In this better understood genres of newscasts and movies, the presence of shots provides cues to "interesting" content. Thus, simple techniques such as preserving the shot boundaries worked reasonably well in such genres to provide an overview. Examples of this approach include [7] and [8], in which the authors use shot-boundary detectors and then use the first/middle/last frames from each shot as the summary. Similarly, Zhang *et al.* [9] choose the first frame of each shot as a keyframe, and then, if the difference between the

Fig. 1. Examples of some unconstrained videos. Each row here shows six frames from videos representing various domains. (a) Cheerleading YouTube video. (b) "Office Tour" YouTube video. (c) Aerial video.

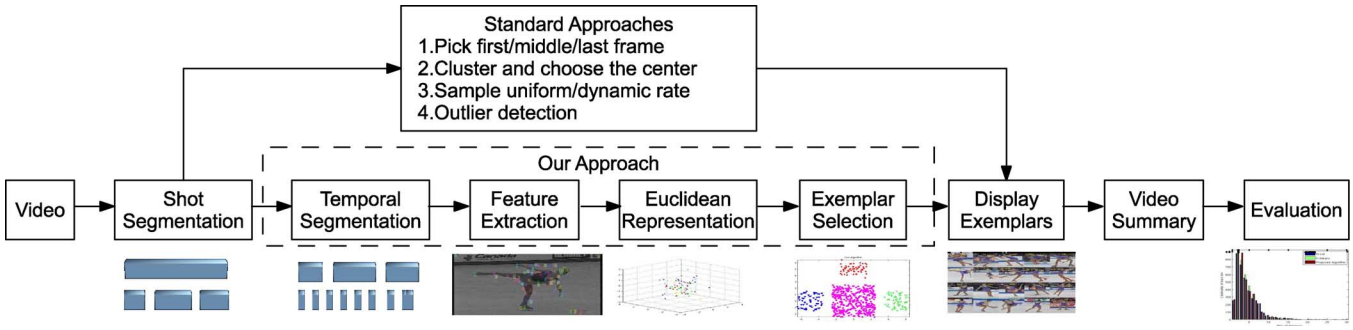


Fig. 2. Overview of existing approaches to video summarization. Existing approaches rely predominantly on computing and processing shots and end up picking exemplars from the dominant event, leaving the interesting but infrequent events.

subsequent frames and the latest keyframe exceeds a threshold, the current frame is also chosen as a keyframe.

However, in unconstrained videos such as the ones shown in Fig. 1, within-shot variations are too large to be ignored. For instance, each of the videos shown in Fig. 1 comes from a *single-shot* video containing 5000 frames. Therefore, a single frame per shot would not be able to capture these variations. So, multiple keyframes/segments must be assigned to each shot based on the dynamics of the shot. Zhang *et al.* [10] and Gunsel *et al.* [11] segment the video into shots and select the first clean frame of each shot as a keyframe. Other frames in the shot that are sufficiently different from the last keyframe are marked as keyframes as well. This shot-based approach captures the original video well when the shots are largely stationary or unchanging. Another approach that was used to solve this problem was to use some form of smart-sampling method that could pick more frames from highly informative regions. Nam and Tewfik [12] generate the summary by adaptive and dynamic sampling of the video sequence. Specifically, they calculate subshot units of the video sequence as well as motion activity for each of them. Subshot units with highly quantized motion-intensity index are then sampled at a higher rate than those with a lower index. Divakaran *et al.* [13] also propose motion-based nonuniform sampling of the frames. The underlying hypothesis of motion-based sampling is that the intensity of motion activity of a video segment is a direct indication of its "summarizability." In [14] and [15], the authors use audio descriptors along with the motion descriptors to summarize the videos. Recently, Chen *et al.* [16] constructed the relational graph using the correspondence between video and text information and then exploited the graph entropy model to detect meaningful shots and relations. De Menthon *et al.* [17] represent the video as a curve in a high-dimensional feature space and use a multidimensional curve-splitting algorithm to linearize the curve and extract the keyframes. The limitation with this class of approaches is that these measures of summarizability are local in nature. In effect, this may lead to important segments being missed while longer segments might show multiple frames with similar content.

To remedy these issues, one needs a principled way to select informative exemplars from all of the available frames or video segments. The simplest of approaches would be to use clustering methods. In this approach, video frames are treated as points in a feature space and then the most representative frames/segments are shown from each cluster as part of the video abstract. Various image-based features like color or motion have been employed for this task. Subsequent to clustering, cluster centers are presented to users as the video summary. This also enables a user to search for relevant clips in the original video by using the clusters as an index. Ferman *et al.* [18] and Zhuang *et al.* [19] use clustering on the frames within each shot. The frame closest to the center of the largest cluster is selected as the keyframe for that shot. This approach does not capture within-shot variations of nonstationary shots well, as only one keyframe was chosen per shot. Hanjalic and Zhang [20] propose another clustering

technique where all video frames are clustered into a variable number of clusters. Cluster validity analysis is then performed to determine the optimal number of clusters n. Similarly, activity specific features are extracted over small segments of videos, and each segment is represented as a collection of features derived from a histogram as in [21] or as a linear dynamical system [22]. The methods presented in [23], [24] have also focused on generating summaries for domain-specific videos where special features using the domain knowledge can be employed. Given a long video, these clustering approaches proceed in an unsupervised fashion to extract summaries.

Although clustering techniques have been quite effective, many of them reduce to choosing representative frames from the most dominant clusters. However, as discussed in [25], in a large class of videos (e.g., surveillance or sports), interesting events happen infrequently among the background of usual events. Many times, these unusual/interesting events are the desired events which are often not captured by traditional clustering approaches. To resolve this issue, another approach was to choose outliers (unusual or uninteresting events) as the exemplars as in [25]. However, unusual events like a commercial break in a soccer video need not be interesting to the end user [25]. This motivates us to propose a summarization method that ensures mutually exclusive exemplars which are also exhaustive.

To complete the discussion, we briefly discuss visualization methods for video summaries. Significant research has also been done on finding novel ways of presenting and visualizing the summary. One of the most commonly used methods to present summaries is via storyboards. Another approach is through mosaicing [26]–[28], which tries to present most of the pixel intensity variations that are spread out over several minutes by piecing together a large mosaic. Another class of approaches collapses the regions of motion into a smaller spatio-temporal volume. In [29], this was done by detecting "tubes" of moving objects in the video and fitting all of the tubes into a coherent video. This approach essentially displays all of the moving objects in the scene with no special regard to what activity is being performed by the object. A content-aware resizing approach is taken in [30], where "seams" of low gradient are successively removed from a video. The resulting video then represents high-gradient information in the video. Another approach for retargeting the image/video data has been presented in [31]. All of these methods are primarily visualization techniques and are best suited for creating visual "thumbnails" of videos. In this work, we restrict ourselves to the problem of choosing "good" exemplars to represent the unconstrained video (of any domain) well and generate a quick overview of the video. Our focus is not on how the summary is visualized, as we simply choose to use one or the other method as appropriate in experiments.

Contributions: We present a mathematical formulation of the video summarization problem as one of maximizing coverage (exhaustive) and diversity (mutually exclusive) and propose an algorithm to solve it. Further, to demonstrate the effectiveness of the proposed criterion, results on four different classes of videos are presented. Subsequently, the summary generated is evaluated using both quantitative measures and user studies.

Organization of the Paper: In Section II, we first discuss the proposed video summarization framework. Section III discusses the role of the proposed criteria in video summarization. Then, in Section IV, we quantify these criteria and provide a mathematical formulation for the summarization problem, followed by a discussion of how to solve the summarization problem in Section V. In Section VI, we propose a few extensions to the proposed cost function. Section VII presents the algorithm for preprocessing and feature extraction from the video. In Section VIII, we provide several experimental results that demonstrate the effectiveness of the proposed approach. In Section IX, we evaluate our method using quantitative and qualitative measures. Finally, discussions and concluding remarks are provided at the end of Section IX.

II. Video Summarization Framework

Before proceeding further, we first discuss the proposed framework for summarizing video content, from which specific algorithms can be derived. Video summarization can be seen as proceeding in a sequence of steps, which are discussed below and are concisely shown in Fig. 2. We also discuss here various considerations at each step.

- **Temporal Segmentation**: A given video first needs to be broken into segments where there is some consistency in viewpoint or scene, for example. This frequently takes the form of shot segmentation. Each shot is further segmented into small video segments. This segmentation should ensure that each of the segments is smooth with minimum variation and consistency of dynamics/view/objects in it. In our implementation, we choose these segments to be of constant length of 20 frames. Choosing such short segments ensures the consistency in each segment.
- **Feature Extraction**: Once coherent shots are identified, we need to extract concise features that capture relevant information from the frames. The video segments are then represented as points in an N-dimensional space. Towards this goal, features are independently extracted from each segment. If the end application requires the summary to capture the content in terms of various spatial appearances and dynamics, features such as the spatio-temporal features suggested in [32] may be chosen. Further, if the property of rate/view/scale invariance is to be incorporated in the summarization system, features with such properties should be chosen. If the summarization system has to use audio cues, features from audio should be extracted [14], [15].
- **Exemplar Selection**: This stage selects important exemplars in the video that capture the essence of the video. Given a set of points in a Euclidean space, the problem of summarization boils down to one of optimally selecting a subset of points (exemplars). At this stage, the measure of optimality becomes crucial. Traditionally, this has been implemented by means of standard clustering approaches like k-means or Ncut. Alternately, optimizing a specified cost function that incorporates desired attributes over the chosen subset [33] can be used for this task. In this work, we adopt the latter approach. We first discuss two criteria "coverage" and "diversity" that a summary should have and then incorporate them into a cost function. Traditionally, as discussed in the related work section, clustering approaches have been adopted for choosing this optimal

subset. Hence, to show the significance of the two criteria, we compare it with the two well-known clustering approaches—k-means and Ncut.

- **Visualization**: The best visualization of this optimal subset (i.e., displaying the exemplars) depends largely on the end application. It can be displayed as a static storyboard or as a temporal concatenation of the chosen video segments. In the latter case additional criteria like "coherence" and "comprehensibility" can be incorporated to make the visualization smoother. However, the focus of this work is to optimally select the exemplars, and we choose here the simplistic way of temporally concatenating the chosen exemplars.

III. Role of Diversity and Coverage in a Précis

Précis is a term that is used to refer to summaries of long documents that encompass the essence of a document without omitting any significant details. This requires the summary to contain the dual properties of diversity and coverage. While easy to state, it is not apparent what this entails in computational terms. Before delving into mathematical details, let us consider what these ideas mean in terms of video summarization. In a large class of videos (e.g., surveillance videos or sports videos), large portions typically contain "uninteresting" events. Thus, standard summarization approaches might miss the really interesting aspects and preserve the dominant uninteresting portions. Most of the clustering-based approaches do not necessarily enforce the criterion that the discovered summary highlight diverse aspects of the video. Further, even if by some preprocessing steps one were to remove the irrelevant parts of a video, the results of optimal subset selection can be easily skewed by uneven statistical distribution of the video segments. An action/event that occurs more frequently than others would skew the chosen subset towards this more dominant event. Thus, motivated by these consideration, we suggest two properties for a good summary of videos: 1) coverage and 2) diversity. The first criterion suggests that the summary should "cover" most of the video in terms of global representation. The second criterion suggests that the elements of the summary should be as distinct as possible.

In this paper, we explore these ideas for the problem of video summarization. To solve this problem, we model the summarization process as searching for a set of video segments from the original video that satisfy the properties discussed above. We provide a conceptually simple quantification of these criteria into a cost function. We show how to find a good summary by a combinatorial optimization approach to minimize the cost function. Experiments show that, even with simple and intuitive metrics as proposed here, significant improvements in summarization quality over traditional approaches are obtained. Further, the problem formulation and solution is largely independent of the choice of features and easily extends to new features and, thereby, to new classes of videos.

It is worthwhile to study the parallel advances in the document mining community which has also addressed similar problems. In fact, the widely used bags-of-words model and topic models have their origins in text-mining and retrieval. One of the important problems in this field is to generate a summary of a single long document by selecting informative sentences [34]. By enforcing that the document summary should contain the properties of "Coverage" and "Orthogonality" [35], the generated summaries were found to be of significantly higher quality. However, the methods proposed in [35] are strongly tied to documents and do not easily generalize to other domains such as videos.

IV. Coverage and Diversity: A Cost Function

Consider a video V as a collection of video segments $V = \{v_1, v_2, \ldots v_n\}$. Let us assume that there exists a representation of each video segment that maps it into a Euclidean representation $\mathbb{F} = \{F_1, F_2 \ldots F_n\}$, where each $F_i \in \mathbb{R}^d$ for some d. This representation could be one of several choices such as bags-of-words [36], 3-D structure tensors [21], or motion history [37]. For manifold-valued representations such as linear dynamic models [38] or covariance matrices [39], we shall assume that we can obtain a Euclidean representation via the logarithmic map on the manifold. Thus, the mapping from video segments to the Euclidean representation can be made flexible, and we do not tie down the method to any specific choice of features.

Once we have made the choice as to what representation is best suited for a class of videos, we can then proceed to quantify and optimize the criteria for summarization. We define a summary S of length K to be a collection of K segment indices $S = \{\alpha_1, \ldots \alpha_K\}$, where $\alpha_i \in \{1, \ldots n\}$ for all i. Now, we consider the set $S_F = \{F_{\alpha_i} | \alpha_i \in S\}$ as the set of summary centroids. The centroids implicitly partition the space of Fs by assigning each F_i to the closest centroid. Let V_{α_i} be the partition corresponding to F_{α_i}, i.e., V_{α_i} are the elements of $\mathbb{F}$, which are closer to F_{α_i} than any other centroid. The *scatter* of V_{α_i} can then be defined as

$$\text{scatter}\,(V_{\alpha_i}) = \sum_{F_k \in V_{\alpha_i}} (F_k - F_{\alpha_i})\,(F_k - F_{\alpha_i})^T. \quad (1)$$

Then, the *coverage* of the summary S can be measured in terms of squared error

$$\text{err}(S) = tr\left[\sum_i \text{scatter}\,(V_{\alpha_i})\right] \quad (2)$$

$$= tr\left[\sum_i \sum_{F_k \in V_{\alpha_i}} (F_k - F_{\alpha_i})\,(F_k - F_{\alpha_i})^T\right]. \quad (3)$$

A high coverage implies a low error, and *vice versa*. On the other hand, the *diversity* of the summary is measured in terms of the scatter of the centroids as follows:

$$\text{div}(S) = tr\left[\sum_i \left(F_{\alpha_i} - \overline{F}\right)\left(F_{\alpha_i} - \overline{F}\right)^T\right] \quad (4)$$

where $\overline{F} = (1/K)\sum_j F_{\alpha_j}$ is the mean of the centroids. From this measure of diversity, we can define the "redundancy" of the summary as $\text{red}(S) = D - \text{div}(S)$, where D is an upper bound on the diversity possible. This measure ensures that $\text{red}(S)$ is always a positive number. However, the precise value of D is not critical for the rest of the discussion. Using, these measures

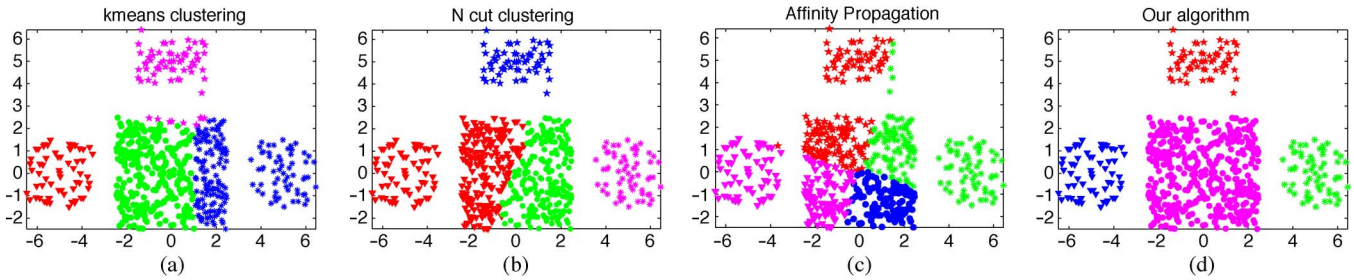


Fig. 3. Comparison of the standard approaches and the proposed cost function in choosing exemplars. Each set is represented in a different color (online version), and the centroid of the set is chosen as the exemplar. Standard approaches do not find "diverse" exemplars. (a) k-means. (b) Ncuts. (c) Affinity propagation. (d) Proposed method.

of coverage and redundancy, the overall cost of a summary is given by

$$J(S) = \alpha * \mathrm{err}(S) + (1-\alpha) * \mathrm{red}(S) \tag{5}$$

with $0 \leq \alpha \leq 1$. The goal of the summarization algorithm is then to minimize the cost-function $\hat{S} = \min_S J(S)$. Here, α represents a weighting parameter that provides relative weightage to each of the criterion. A value of $\alpha = 1$ boils the problem down to a standard least squares problem which can be solved by k-means or k-medoids. We let the value of α be a user-defined parameter whose value should be decreased if the diversity criterion is to be emphasized. Note that the precise value of D does not affect the minimization of the cost function, since it only adds a constant value to the cost. It provides a convenient method to pose the problem as the minimization of a positive function.

Relation With Other Cost Functions: Here, we discuss the relationship between the cost function in (5) with other well-known cost functions. k-means and k-medoids minimize a cost function which is similar to the first part of (5). The difference is the diversity criterion, in the absence of which this subset selection can be easily skewed toward dominant events. The diversity criterion ensures that even infrequent but interesting events are represented in the summary. Affinity propagation [40] is a relatively new algorithm which also minimizes the sum of squares error criterion, but does not require initialization. Normalized-cuts (Ncuts) [41] is a graph-partitioning algorithm which searches for subgraphs with large "internal association" normalized by its association with the rest of the graph. The internal association can be interpreted as being similar to the coverage criterion. But the normalization term in Ncuts forces the algorithm to find balanced clusters. This does not ensure that the cluster centroids highlight diverse aspects of the video.

The diversity criterion looks similar in form to the "between-class" scatter criterion, and coverage is similar in form to the "within-class" scatter used in Fisher's linear discriminant analysis (LDA), but differs from it in several significant ways. First, there are no "class labels" in the current scenario. Second, we are not interested in searching for projections of the data, as is the case in LDA. Further, if maximizing diversity ($\alpha = 0$) were the only criterion, then the proposed method would be sensitive to outliers. However, the coverage criterion prevents the proposed cost function from latching on to outliers in the data.

We illustrate these ideas using a synthetic experiment for optimal subset selection from a set of points in $\mathbb{R}^2$. For this experiment, we synthetically generated points in $\mathbb{R}^2$ in four distant sets. One of the sets was deliberately made larger than others. Then, we used standard approaches, such as k-means, Ncuts, and affinity propagation, to choose four exemplars. We show the results in Fig. 3(a)–(c). As can be seen, due to the dominance of the central set, standard cost functions pick multiple exemplars (centroid) from the same set. On the other hand, the results of exemplar selection using the proposed cost function are shown in Fig. 3(d). It can be seen that the proposed cost function finds a diverse set of exemplars.

V. Optimizing the Cost Function

Here, we discuss how to optimize the cost function $J(S)$ described in Section IV. First, we observe that this is a combinatorial optimization problem where we need to find optimal solutions (not necessarily unique) in the (discrete) problem space. To find an optimal solution in a combinatorial optimization problem, the difficulty arises from the fact that there is a lattice of feasible points. In some of the other optimization problems like linear programming, this is not the case, as the feasible region is a convex set. In fact, if we replace the least squares cost function of the classical k-medoids partitioning problem with the proposed cost function, then we note that both problems are similar in complexity. As the k-medoids problem is known to be NP-hard, the current problem is also NP-hard. This motivates the need to solve this problem by an approximate method. Observing the similarity between the current problem and classical k-medoids partitioning problem, we adopt the iterative procedure discussed in Algorithm 1.

Algorithm 1: Algorithm for minimizing the proposed cost function

1.1. Start with a random initialization of k centroids $S_F = \{S_{F_1}, S_{F_2}, \ldots, S_{F_k}\}$.
1.2. Assign each of the remaining points in $\mathbb{F}$ to the nearest point in S_F.
1.3. Randomly select a point $S_{F_{\mathrm{rand}}} \in \mathbb{F} \setminus S_F$.
1.4. Compute the difference $J_{\mathrm{swap}} = J_{\mathrm{new}} - J_{\mathrm{old}}$ that results by swapping $S_{F_{\mathrm{rand}}}$ with S_{F_j} where j is a randomly chosen index.
1.5. If $J_{\mathrm{swap}} < 0$, replace S_{F_j} with $S_{F_{rand}}$. Else retain the original S_F.
1.6. Repeat steps 1.2–1.5 till convergence or if maximum iterations are exceeded.

We note that, at each iteration, the algorithm chooses a solution that either decreases the cost or leaves it unchanged. Hence, the optimization always proceeds in a "greedy" fashion. Thus, it is likely that the solution may end up in a local minima. The next question to address is how many iterations are needed before the algorithm converges. We do not have a clear answer for this question, but it was observed that, as the number of points increases, the iterations required to converge increase exponentially. Further, due to the "greedy" nature, the final solution depends on the initialization. Thus, to address these issues, we experimented with several random initializations. For each initialization, we let the algorithm run for a large number of iterations ($\sim 10^3$). From the set of solutions thus obtained, we choose the one with the minimum cost. We found that this strategy works well in most cases.

Computational Complexity: If n is the number of data-points, k is the number of exemplars to be selected and t is the number of iterations, then the computational complexity of algorithm is $O(k(n-k)t)$. Since $n \gg k$, the complexity is linear all three variables.

VI. Weighted Kernel Coverage and Diversity

In many cases, features are usually not linearly separable in the feature space, but may require nonlinear separating boundaries. In the existing literature, this problem is commonly solved in a reproducing kernel Hilbert space (RKHS) instead of in the original feature space such as in [42]. The basic assumption is that there exists a mapping into a high-dimensional RKHS, where linear separability is possible. The problem is implicitly solved in RKHS via a suitably defined Mercer kernel. Here, we shall discuss a kernel extension of the proposed cost function that enables us to optimize diversity and coverage in the RKHS.

Let $\phi_i = \phi(F_i)$ be the mapping of $F_i \in \mathbb{R}^d$ to the higher dimensional RKHS space $\mathcal{H}$. Let P_{α_i} be the partition of the points in the $\mathcal{H}$ space as

$$err(S) = \left[\sum_i \sum_{\phi_k \in P_{\alpha_i}} w_k \left\| \phi_k - \phi_{c_i} \right\|^2 \right] \tag{6}$$

where w_k are nonnegative weights and c_i represents the index of the centroid of the i^{th} partition (P_{α_i}) in the $\mathcal{H}$ space as

$$\begin{aligned} c_i &= \arg\min_{c_i} \left[\sum_{\phi_k \in P_{\alpha_i}} w_k \left\| \phi_k - \phi_{c_i} \right\|^2 \right] \; s.t.\; \phi_{c_i} \in P_{\alpha_i} \\ &= \arg\min_{c_i} \left[\sum_{\phi_k \in P_{\alpha_i}} \left(w_k \phi_k^T \phi_k + w_k \phi_{c_i}^T \phi_{c_i} - 2 w_k \phi_k^T \phi_{c_i} \right) \right] \\ &= \arg\min_{c_i} \left[\sum_{\phi_k \in P_{\alpha_i}} \left(w_k K_{kk} + w_k K_{c_i c_i} - 2 w_k K_{k c_i} \right) \right] \end{aligned}$$

where K is the kernel Gram matrix with $K_{ij} = \phi_i^T \phi_j$ to yield

$$\begin{aligned} \mathrm{err}(S) &= \left[\sum_i \sum_{\phi_k \in P_{\alpha_i}} \left(w_k \phi_k^T \phi_k + w_k \phi_{c_i}^T \phi_{c_i} - 2 w_k \phi_k^T \phi_{c_i} \right) \right] \\ &= \left[\sum_i \sum_{\phi_k \in P_{\alpha_i}} \left(w_k K_{kk} + w_k K_{c_i c_i} - 2 w_k K_{k c_i} \right) \right]. \end{aligned}$$

Fig. 4. Example frames from figure skating [43] video sequence. Spatio-temporal patches (extracted using [32]) are shown overlaid on the images.

The *diversity* of the summary in the $\mathcal{H}$ space is measured similarly in terms of the scatter of the centroids as follows:

$$\mathrm{div}(S) = \left[\sum_i w_{c_i} \left\| \phi(c_i) - \overline{p_c} \right\|^2 \right]$$

where

$$\begin{aligned} \overline{p_c} &= \frac{\sum_i w_{c_i} \phi_{c_i}}{\sum_i w_{c_i}} \\ &= \left[\sum_i w_{c_i} \left(K_{c_i c_i} - \frac{2 \sum_j w_{c_j} K_{c_i c_j}}{\sum_j w_{c_j}} + \frac{\sum_{c_j, c_l} K_{c_j c_l}}{\left(\sum_j w_{c_j} \right)^2} \right) \right]. \end{aligned}$$

Then, defining the redundancy of the summary as earlier $\mathrm{red}(S) = D - \mathrm{div}(S)$, the overall cost function $J(S)$ is given by

$$J(S) = \alpha * \mathrm{err}(S) + (1-\alpha) * \mathrm{red}(S). \tag{7}$$

As can be seen, the mapping ϕ need not be computed explicitly, as the resulting cost function and its optimization only require the knowledge of dot-products in the RKHS. These dot-products can be evaluated by using a Mercer kernel. This kernel extension can be used when no separating hyperplane exists in the feature space and a "better" linear partition can be found in a high-dimensional space. The weights can be uniform if no prior knowledge/preferences exist about the summary of the given video. In a relevance feedback scenario, the user can choose more relevant exemplars from the initial set of returned exemplars, which can be used to refine the summary using the above formulation by assigning higher weights to the chosen exemplars. The experiments presented in this paper simply use $K_{ij} = x_i^T x_j$. This worked well in all of the experiments. However, the results derived in this section provide a principled means to optimize the performance to even higher levels. However, we do not explore this avenue further in the experiments.

VII. Video Analysis and Feature Extraction

Here, we discuss the preprocessing of the video content and how we obtain the Euclidean representation of each video segment. Before we discuss the specific choices made, we would like to emphasize that the optimization cost function and algorithm for exemplar selection are independent of the choices made in lower level modules.

A. Feature Extraction and Euclidean Representation

In the first step, we perform video segmentation or shot segmentation, i.e., partition the video sequence into shots. Once the shots are detected, we extract spatio-temporal features for the Euclidean representation of the video. For this, we adopt the

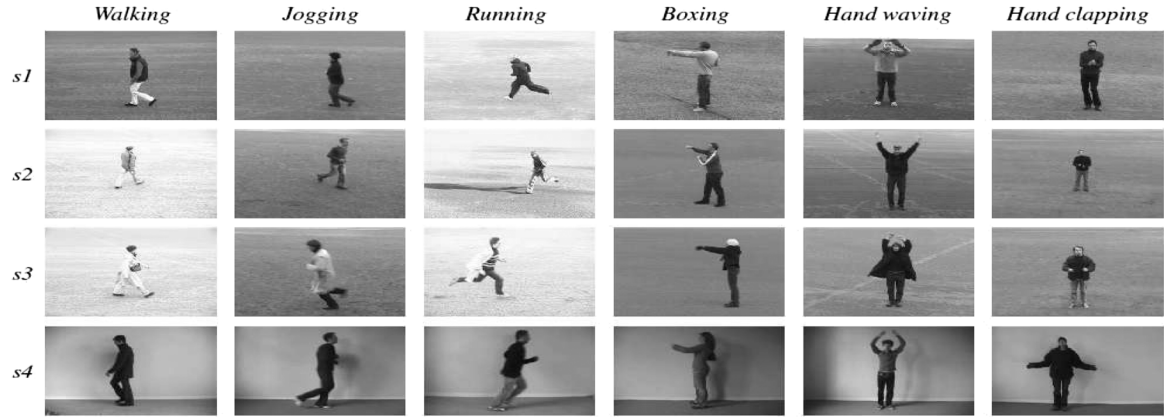


Fig. 5. Example frames from video sequences in the KTH dataset [44]. Figure reproduced from [44].

Fig. 6. Summary of KTH human action database video. Summary of length $K = 6$. (a) K-means. (b) Ncut. (c) Proposed algorithm.

widely used bag-of-words approach to represent the video segments. Each video segment is assumed to be 20 frames long (though a multiscale version is possible here). In the bag-of-words formalism, a long video is considered to be a "document" or as a collection of sentences, where each segment corresponds to a "sentence." In turn, each sentence is considered as a vector in an N-dimensional Euclidean space with the ith value as the indicator of the ith word of the codebook. The words are learned by clustering descriptors extracted around interest points from the entire long video. To extract the interest points, we used the method proposed in [32]. These interest points are the local maxima of the response function $R = (I * g * h_{\mathrm{ev}})^2 + (I * g * h_{\mathrm{od}})^2$ extracted from a sequence of images. Here, I is the sequence of images, $g(x, y, \sigma)$ is the 2-D Gaussian smoothing kernel applied along the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1-D Gabor filters applied temporally. The extracted cuboids are further processed by simply flattening them as suggested in [32]. The detector parameters are set to $\sigma = 2$ and $\tau = 2.5$. As described in [36], we build our codebook by clustering these flattened cuboids into $N = 500$ clusters. The bag of words approach is scalable to long video sequences. Example of these extracted features have been shown in Fig. 4. These features have been overlaid on the corresponding frames and shown.

Another question that remains to be answered is how do we decide the number of exemplars that should be chosen per shot. Any appropriate measure on the length of the summary would be on the whole video rather than each shot. Intuitively, a shot with higher variation should be represented with more exemplars than a shot with lesser variation. So, the number of key video segments chosen per shot is decided using the covariance matrix of the shot. The motivation behind using the covariance matrix is because it captures the "variation" within a shot well and, hence, can be used to assign the number of key video segments. The covariance matrix is calculated for each shot, denoted by $C(\mathrm{shot}_i)$, using the Euclidean representation of each video segment in the shot. This matrix is then used to calculate the number of exemplars per shot as

$$K_{\mathrm{shot}}(i) = \left\lceil \frac{tr\left(C(\mathrm{shot}_i)\right)}{\sum_i tr\left(C(\mathrm{shot}_i)\right)} * K_{\mathrm{video}} \right\rceil \tag{8}$$

where $tr(C(\mathrm{shot}_i))$ represents the trace of the covariance matrix of the shot. $K_{\mathrm{shot}}(i)$ is the number of exemplars for the ith shot while K_{video} is the number of exemplars for the video. $\lceil . \rceil$ represents the standard ceiling function.

VIII. Experiments

Here, we describe experiments demonstrating video summarization using the proposed algorithm and a few illustrative comparisons. We provide experimental results on four representative videos: 1) constrained KTH human action dataset [44]; 2) unconstrained figure skating dataset [43]; 3) unconstrained aerial surveillance VIVID dataset; and 4) an office-tour video from YouTube.[1]

[1]Video results are also available at http://www.umiacs.umd.edu/users/nshroff/Precis.html.

Fig. 7. Figure skating video summary ($K = 20$). Both k-means and Ncut selects around 11 exemplars from "glide" event, while the proposed algorithm picks more diverse and "interesting" exemplars. (a) Figure skating video summary using the k-means algorithm. (b) Figure skating video summary using the Ncut algorithm. (c) Figure skating video summary using the proposed algorithm.

A. *Constrained KTH Action Database*

The first experiment was performed on the publicly available KTH human action dataset [44]. The dataset consists of six human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 subjects in four different scenarios (backgrounds), as shown in Fig. 5. In our experiment, we constructed a long video by appending a subset of this dataset which includes the first five subjects performing each of the six actions in the scenario $s4$.

The set of vectors was constructed as discussed in Section VII-A, and is then fed to the optimization algorithm discussed in Section V. The value of $\alpha = 0.45$ was used. Since there are precisely six distinct activities in the whole video, we generated a summary of length $K = 6$ to test the strength of the proposed framework. Ideally, each element in the generated

Fig. 8. (a) VIVID video. Frame of this 7200 frames long video has been mosaicked into three parts. These mosaics are shown to give the readers an idea about the full content of the video. (b) Summary generated using Ncut ($K = 15$ length). Best viewed with pdf magnification.

summary should represent a distinct activity. The results of the summarization are shown in Fig. 6(c). Each row represents a video segment of the generated summary. Due to space constraints, we only show six uniformly sampled frames from the 20 frames of each video segment. From the available ground truth, we see that the elements of the summary correspond to the following activities—Boxing, Boxing, Running, Walking, Hand-clapping, Hand-waving—in the order they appear in Fig. 6(c). We see that five of the six activities are captured in the summary. The only activity missing is Jogging. It is important to note that Jogging is very similar to Running and Walking. Hence, the summarization algorithm did not capture Jogging as a distinct activity. For comparison, we show the results of exemplar selection with k-means and Ncut in Fig. 6(a) and (b). For the k-means algorithm, we ran the algorithm a few times and chose the clustering that had the lowest clustering cost.

B. Unconstrained SFU Figure Skating Video

In the next experiment, we used the skating videos from [43], which are completely unconstrained videos with real pan, tilt, and zoom of the camera with rapid motion of the skater. These figure skating videos consist of a few established elements or moves such as jumps, spins, lifts and turns. A typical performance by a skater or a pair of skaters includes several of these elements each performed several times. We show our results on a single skater video which consists of about 3500 frames. The vector representation of each segment is obtained as discussed in Section VII-A. The summary vectors are then chosen using the k-means, Ncut, and the proposed algorithm (with $\alpha = 0.45$). Representative frames from each video segment of the generated summary are shown in Fig. 7(a)–(c), respectively. Each frame has been marked with the name of the figure skating element being performed in the corresponding video segment. Note that several segments of the k-means and Ncut summary correspond to "gliding" [Figs. 7(a) and (b)], and it misses out on some of the significant spins, jumps and spirals in the video. On the contrary, the diversity criterion of the proposed algorithm ensures that these varied actions are captured in the summary, as shown in Fig. 7(c).

C. VIVID Surveillance Video

Here, we show the summarization results on DARPA's VIVID Dataset. The videos are low-resolution aerial videos captured by an unmanned aerial vehicle (UAV). We choose a video of around 7200 frames. To show the complete content of the original video, it has been mosaicked into three parts and shown in Fig. 8(a). Videos in this dataset are continuously captured from a single camera and, hence, form a single-shot video. Vectorial representation of each video segment (20 frames) is obtained as described in Section VII-A. $K = 15$

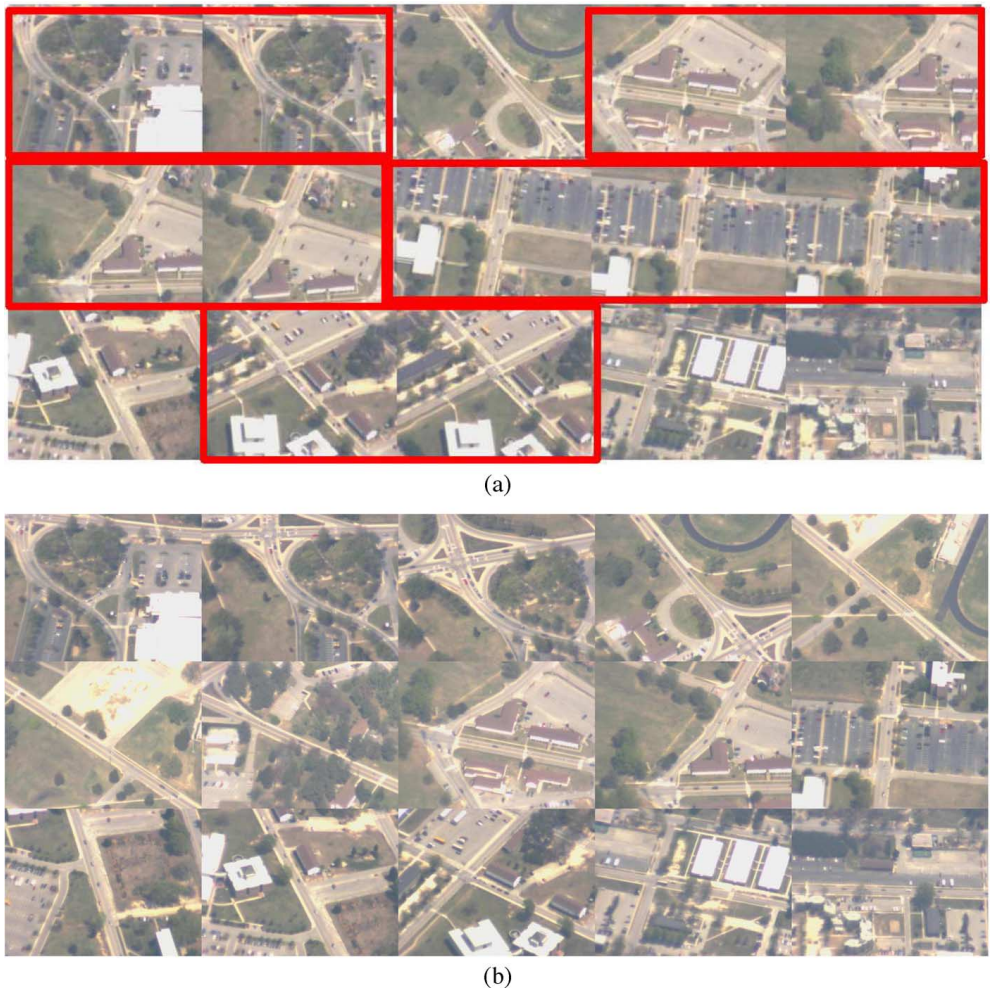


Fig. 9. VIVID video summary ($K = 15$ length). Shown are the central frames from each video segment of the summary generated. Comparatively, lot many redundant exemplars are chosen by k-means and Ncut. (a) VIVID video summary using the k-means algorithm. (b) VIVID video summary using the proposed algorithm.

exemplars are chosen as the summary using Ncut, k-means, and the proposed algorithm (with $\alpha = 0.45$). As is evident from the results shown in Figs. 8(b) and 9(a) and (b), the summary generated by the proposed algorithm highlights diverse aspects which Ncut or k-means algorithm completely miss out on.

D. Unconstrained YouTube Video "Office Tour"

In the next experiment, we downloaded a 5-min homemade video "Office Tour" from YouTube. As indicated by the name, this video is about a man touring an office and meeting various employees. Since the video was captured by a handheld camera, it has jitter and other motion artifacts caused due to the hand movements. This video was captured in one long shot and had around 7500 frames. This video was chosen to demonstrate the strength of the proposed approach. Vectorial representation of each video segment (20 frames) is obtained as described in Section VII-A. Seventy-two uniformly sampled frames from the original video are shown in Fig. 10(a). A summary of 20 exemplars was generated using each of the algorithms: (a) Ncut, (b) k-means, and (c) the proposed algorithm. Central frames from each of the exemplar segments are shown in Figs. 10(b) and 11(a) and (b), respectively.

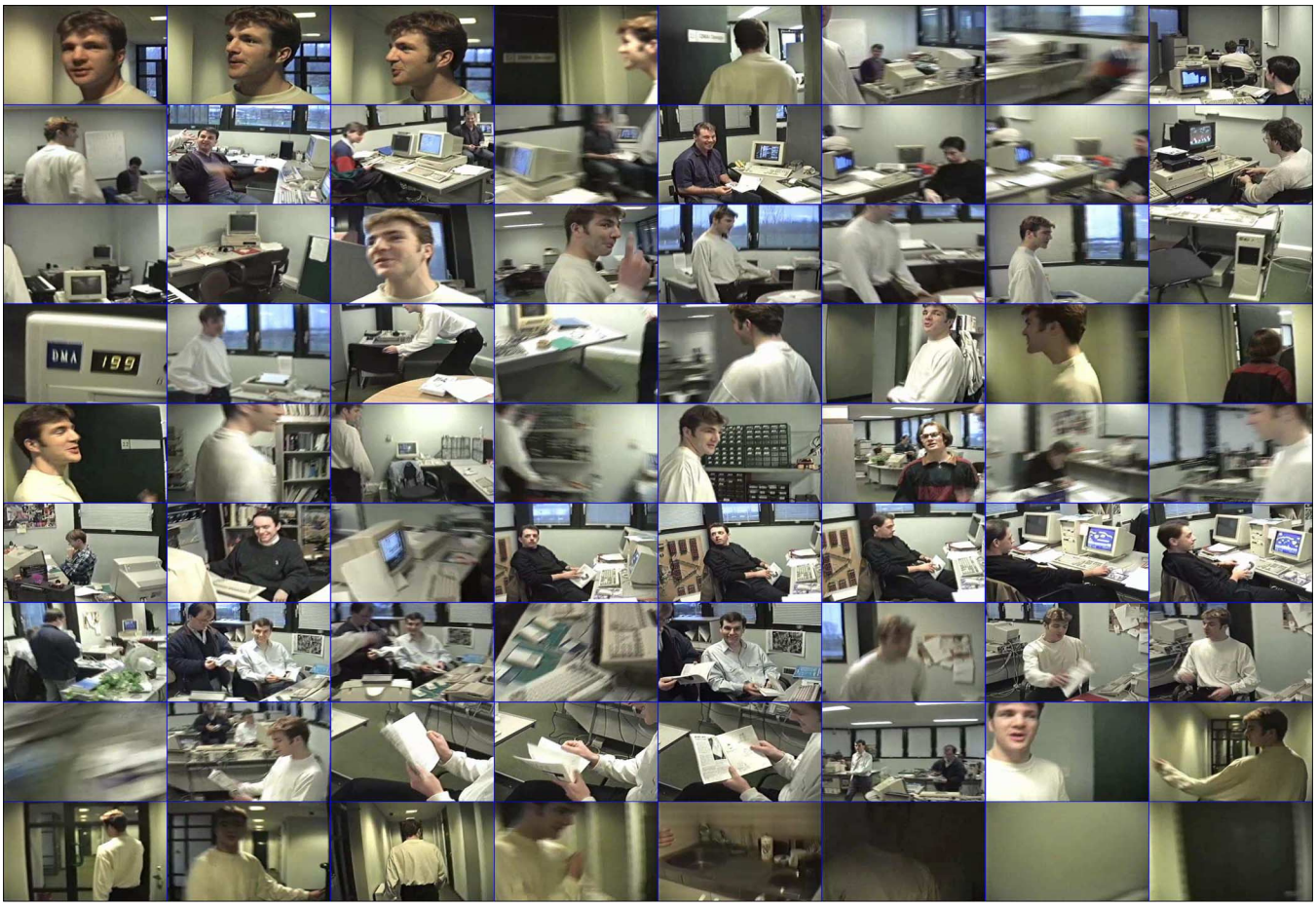
Visualizing the Increase in Diversity: To visualize the increase in diversity, we calculate the pairwise distance between the centroids for each of the three unconstrained videos used in our experiments. Larger pairwise distances signify that the centroids are more spread out, leading to greater diversity of the summary. In Fig. 12(a)–(c), we show the histograms of pairwise distances between the centroids for (a) figure skating, (b) VIVID video, and (c) YouTube video, respectively. As is evident, the proposed algorithm has a larger diversity in all three videos.

IX. Evaluation

A. Qualitative Evaluation

Here, we show results of a user study that we conducted to measure the effectiveness of the proposed approach to summarization. Such a study helps in understanding the factors relevant to human perception. We now describe the user study that we conducted.

Twelve voluntary subjects took part in the evaluation. None was involved in the design/implementation of the proposed algorithm and, thus, were unaware of what to expect. The sub-

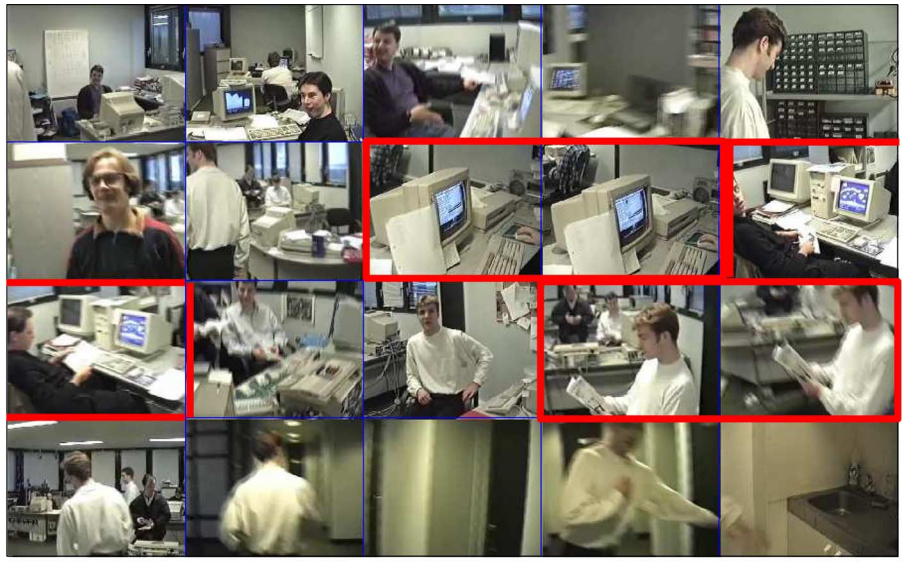


(a)

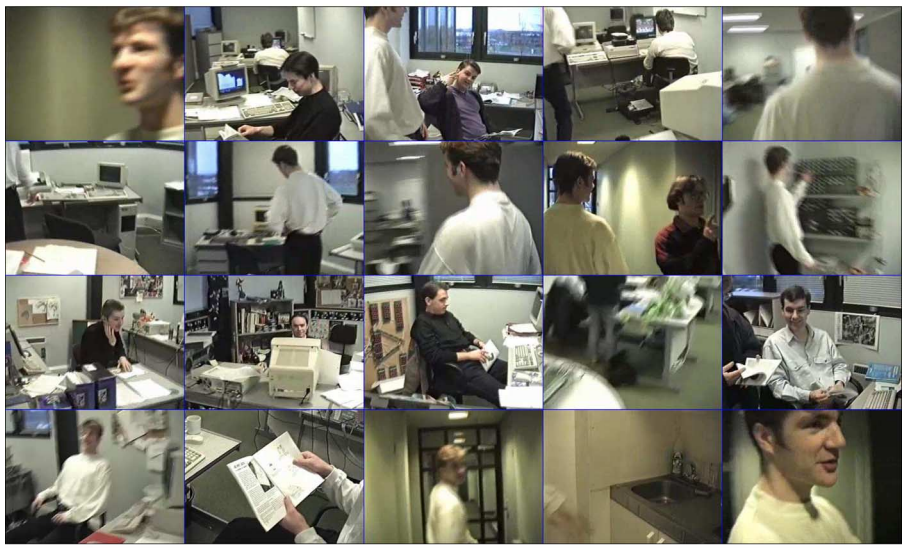
(b)

Fig. 10. YouTube video "Office Tour." This video is approximately 5 min (7500 frames) long. (a) Seventy-two uniformly sampled videos. (b) $K = 20$ length summary generated using Ncut. Shown are the central frames of each video segment in summary. Redundant frames are marked in red boxes (online version). (a) YouTube video uniform samples. (b) YouTube video summary using the Ncut algorithm.

jects were allowed to watch the video sequences—both the originals and the summaries—as many times as they desired in any order they wanted. Breaks were allowed, and the subjects could change any of the answers whenever they wanted. First the original video (in the form of full-length video or large number of uniform samples) were shown to the subjects. Then each of the summaries (generated by proposed algorithm, K-means and Ncut—denoted as Algorithms A/B/C) were shown. The algorithm used to generate the summary was *not* revealed to the user but simply labeled as A/B/C. After showing each set of video



(a)



(b)

Fig. 11. YouTube video summary ($K = 20$). Shown are central frames of each video segment of summary. Redundant frames are marked in red boxes. (a) YouTube video summary using the k-means algorithm. (b) YouTube video summary using the proposed algorithm.

TABLE I
PARAMETERS CHOSEN FOR EACH EXPERIMENT. BOW HERE IMPLIES BAG-OF-WORDS DISCUSSED IN SECTION VII-A

Video	Total frames	Summary Length K	Weight α	# frames per video segment	BoW τ	BoW σ	BoW N
KTH	14400	6	0.45	20	2	2.5	500
Figure Skating	3819	20	0.45	20	2	2.5	500
VIVID video	7200	15	0.45	20	2	2.5	500
Youtube video	7500	20	0.45	20	2	2.5	500

and the corresponding summaries, users were asked to fill in an evaluation form with the questions and possible answers shown in Table II.

After the evaluations were completed by all of the subjects, they were compiled and a few statistics were computed. The results are shown in Fig. 13(a)–(c). The first two figures correspond to questions about the comparative performance between the three algorithms. As can be seen in Fig. 13(a), the proposed algorithm was given the highest score in all four questions and, hence, demonstrates superior quality of the summary. In Fig. 13(b), as expected, the proposed algorithm has the least number of redundant keyframes in its summary and the least number of important segments left out. Fig. 13(c) focuses on questions that ask whether the user wants to watch the full-length video after watching the summaries with possible answers of Yes/No/Maybe. Around 80% of users agreed that the summary could help them in deciding if it was interesting to them or not. The subjective answers to survey question number 8 highlighted that the redundancy was reduced by the proposed algorithm and therefore more amount of information was being

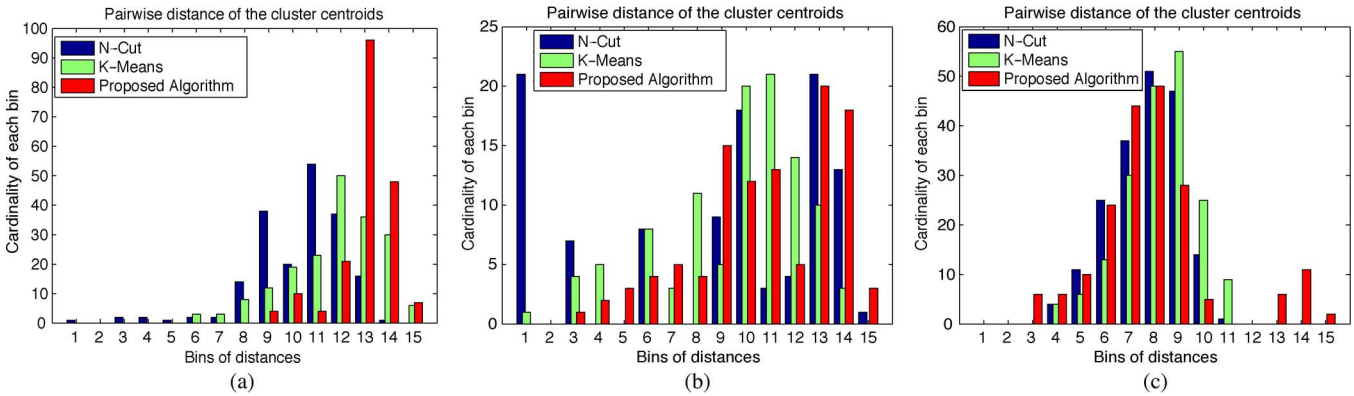


Fig. 12. Histogram of pairwise distances of the centroids. Note that the proposed algorithm (red bars, online version) has a greater number of points in the higher distance bins which implies that the centroids chosen are more distant than the Ncut (blue bars, online version) and k-means (green bars, online version). (a) Figure skating. (b) VIVID video. (c) YouTube video.

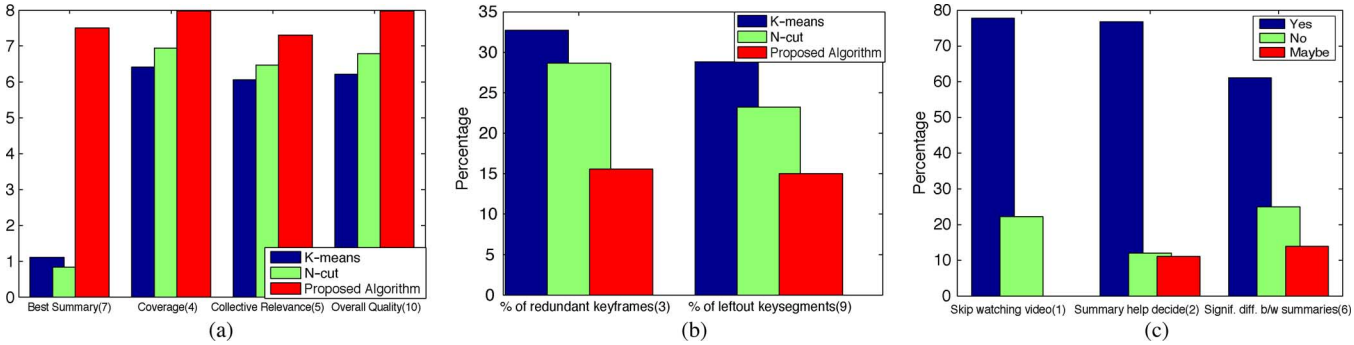


Fig. 13. Responses for (a) survey questions 7, 4, 5, and 10, (b) survey questions 3 and 9, and (c) survey questions 1, 2, and 6. Numbers in the bracket indicates the corresponding survey question in Table II.

TABLE II
QUESTIONS USED ALONG WITH THEIR POSSIBLE ANSWERS IN THE QUALITATIVE EVALUATION OF THE SUMMARY

S.No.	Questions	Possible Answer
1	Can you skip watching the video as you already know almost all of the content of the video?	Yes / No
2	Did the summary help you decide whether you would like to watch the original video or not?	Yes / No / May be
3	How many keyframes in each summary were redundant? (Give an approximate percentage).	0 - 100 %
4	Did it capture the essence of the video? i.e., rate the coverage of the video.	1 - 10 (10 is best)
5	Rate the collective relevance of the individual keyframes in each summary.	1 - 10
6	Are there significant differences between the various automated summarization methods?	Yes / No / May be
7	Choose the best summarization scheme.	A / B / C
8	Why do you find it better than other sequences?	Subjective
9	How many important segments do you think were left out? (Give an approximate percentage).	0 - 100 %
10	Overall quality of the summary generated by each scheme.	1 - 10
11	Comments (if any)	Subjective

captured in the summary. This user-based evaluation further supports our conclusions.

B. Quantitative Evaluation

We evaluate the summary further, this time quantitatively using a reconstruction error-based cost function. Here, we would like to measure how well the summary centroids can reconstruct the video. To measure this, we define $P(S)$ as the count of the frames whose reconstruction error using the exemplars is above the threshold γ

$$P(S) = \sum_i u\left[\left(min_w \|F_i - Sw\|^2\right) - \gamma\right] \tag{9}$$

where $S_{d\times K}$ is the matrix of K exemplars each with Euclidean representation in $\mathbb{R}^d$. Here, $u[n]$ is the unit step function, F_i is the d-dimensional Euclidean representation of the ith frame, and γ is the threshold of the reconstruction error above which the frame is counted to be lying in the null space of S. $w_{K\times 1}$ is the vector of the weights assigned to each exemplar. The reconstruction error for each frame is minimized over the w for each F_i, which is then chosen to be $w = S^+F_i$, where S^+ is the pseudo-inverse of S.

This cost function supports the intuition about the basic feature of a "good" summary that all of the frames of the video must lie in the space spanned by the linear combination of the exemplars. Therefore, the fewer the number of frames in the null space of the exemplars, the better the summary.

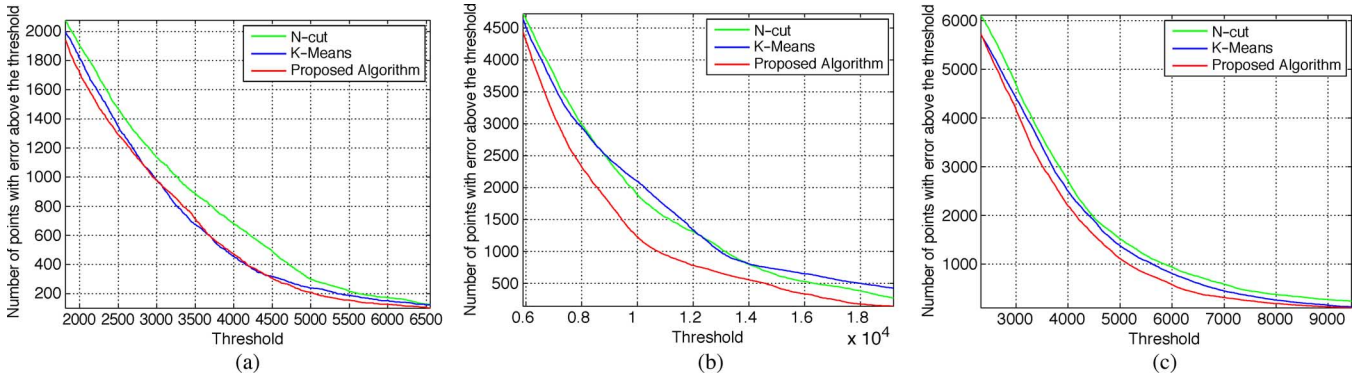


Fig. 14. Reconstruction error-based objective evaluation and comparison of the (i) Ncut (green, online version), (ii) k-means (blue, online version), and (iii) Proposed algorithm (red, online version). (a) Figure skating. (b) VIVID video. (c) YouTube video.

TABLE III
RECONSTRUCTION ERROR EVALUATION. SHOWN HERE ARE THE NUMBER OF FRAMES IN THE NULL SPACE OF THE EXEMPLARS CHOSEN BY DIFFERENT ALGORITHMS FOR THE SPECIFIED THRESHOLD. THE PROPOSED ALGORITHM HAS THE LEAST NUMBER OF FRAMES IN THE NULL SPACE AND, HENCE, HAS MUCH MORE DIVERSE EXEMPLARS CHOSEN

Video	Total frames	Threshold	N-cut	k-means	Proposed Algorithm
Skating	3819	4.6e3	443	303	281
VIVID	7200	1.7e4	456	573	248
Youtube	7500	8492	325	201	154

Any evaluation based on the proposed exemplar selection cost function $J(S)$ would have been very much tuned to the choice of features. Hence, the choice of the evaluation cost function $P(S)$ can give importance to the visual aspect of the summary. As this quantitative evaluation $P(S)$ is independent of the proposed approach to summarization, it makes it usable for evaluation of any summarization algorithm in general.

For the reconstruction, each frame was considered as a gray-level intensity image. A histogram of the gray values of each frame and the exemplars was used for the evaluation. This representation further brings out the independence of the evaluation function to the summarization technique method proposed above.

The results of this evaluation are shown in Fig. 14. Shown in this plot is the number of frames in the null space of the exemplars chosen by different algorithms with varying threshold. Table III shows the number of frames in the null space at one particular threshold. This evaluation points out that the proposed algorithm has the least number of frames in the null space. The figure skating video has a lot more abrupt changes and dynamics in it as compared to the relatively smoothly changing VIVID video or YouTube video. Therefore, the ratio of frames in the null space is higher as compared to the other two videos. This quantitative evaluation further supports the user-based evaluation in emphasizing the role of the diversity criterion in the video summarization.

C. Discussions and Future Work

We have proposed an unsupervised method for summarizing long videos by quantifying two important criteria, namely, coverage and diversity. We have shown that fairly simple definitions of these concepts translate to significant improvements in summarization quality over more traditional approaches. The improvement has been demonstrated through experiments on four different class of videos. The summaries thus produced have been evaluated both qualitatively (user-based evaluation) and quantitatively (reconstruction error function).

The goal of Video Précis is to provide a condensed and succinct representations of the content of a video. But defining which video segments are "interesting" is a very subjective process. It is also very difficult to map human cognitive abilities into an automated abstraction process. The difficulty of the problem increases as the properties of a video summary also depend on the application domain, the characteristics of the sequences to be summarized, and the purpose of the summary. We have proposed a method that tries to optimize between the conflicting requirements of coverage and diversity and shown that it is well suited to summarize a large class of unconstrained videos.

ACKNOWLEDGMENT

The authors would like to thank Dr. A. C. Sankaranarayanan, Dr. A. Veeraraghavan, and other members of Computer Vision Laboratory, University of Maryland, for helpful discussions. The authors would also like to thank Dr. M. Ramachandran for providing us with the mosaicing code. The authors also thank the anonymous reviewers for their suggestions in improving the manuscript.

REFERENCES

[1] B. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., Applic.*, vol. 3, no. 1, pp. 1–37, Feb. 2007.
[2] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques," HP Laboratory/HPL-2001-191 Tech. Rep., Jul. 2001.
[3] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Representation*, vol. 19, no. 2, pp. 121–143, Feb. 2008.
[4] X. Zhu, A. Elmagarmid, X. Xue, L. Wu, and A. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.
[5] J. Oh, Q. Wen, S. Hwang, and J. Lee, "Video abstraction," in *Video Data Management and Information Retrieval*, S. Deb, Ed. Hershey, PA: Idea Group Inc. and IRM Press, 2004, pp. 321–346.
[6] C. Taskiran and E. Delp, "Video summarization," in *Digital Image Sequence Processing, Compression, and Analysis*, T. Reed, Ed. Boca Raton, FL: CRC, 2005, pp. 215–231.

[7] B. Shahraray and D. Gibbon, "Automatic generation of pictorial transcripts of video programs," in *Proc. Soc. Photo-Opt. Instrum. Eng.*, San Jose, CA, Feb. 1995, vol. 2417, pp. 512–518.
[8] S. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia Mag.*, vol. 1, no. 2, pp. 62–72, Summer, 1994.
[9] H. Zhang, J. Wu, D. Zhong, and S. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recogn.*, vol. 30, no. 4, pp. 643–658, Apr. 1997.
[10] H. Zhang, C. Low, S. Smoliar, and J. Wu, "Video parsing, retrieval and browsing: An integrated and content-based solution," in *Proc. ACM Int. Conf. Multimedia*, San Francisco, CA, Nov. 1995, pp. 15–24.
[11] B. Gunsel, Y. Fu, and A. Tekalp, "Hierarchical temporal video segmentation and content characterization," in *Proc. Soc. Photo-Opt. Instrum. Eng.*, Dallas, TX, Nov. 1997, vol. 3229, pp. 46–56.
[12] J. Nam and A. Tewfik, "Video abstract of video," in *Proc. IEEE 3rd Workshop Multimedia Signal Process.*, Copenhagen, Denmark, Sep. 1999, pp. 117–122.
[13] A. Divakaran, R. Radhakrishnan, and K. Peker, "Video summarization using descriptors of motion activity: A motion activity based approach to key-frame extraction from video shots," *J. Electron. Imaging*, vol. 10, no. 4, pp. 909–916, Oct. 2001.
[14] A. Divakaran, K. Peker, R. Radhakrishnan, Z. Xiong, and R. Cabasson, "Video summarization using mpeg-7 motion activity and audio descriptors," in *Video Mining*, A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds. Boston, MA: Kluwer, Oct. 2003, p. 91.
[15] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. IEEE Int. Conf. Image Process.*, Barcelona, Spain, Sep. 2003, vol. 1, pp. 5–8.
[16] B. Chen, J. Wang, and J. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 295–312, Feb. 2009.
[17] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *Proc. ACM Int. Conf. Multimedia*, Bristol, U.K., Sep. 1998, pp. 211–218.
[18] A. Ferman and A. Tekalp, "Multiscale content extraction and representation for video indexing," in *Proc. Soc. Photo-Opt. Instrum. Eng.*, Dallas, TX, Nov. 1997, vol. 3229, pp. 23–31.
[19] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Process.*, Chicago, IL, Oct. 1998, vol. 1, pp. 866–870.
[20] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1280–1289, Dec. 1999.
[21] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Kauai, HI, Dec. 2001, vol. 2, pp. 123–130.
[22] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Unsupervised view and rate invariant clustering of video sequences," *Comput. Vis. Image Understanding*, vol. 113, no. 3, pp. 353–371, Mar. 2009.
[23] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recogn. Lett.*, vol. 25, no. 7, pp. 767–775, May 2004.
[24] D. Zhong, R. Kumar, and S. Chang, "Real-time personalized sports video filtering and summarization," in *Proc. ACM Int. Conf. Multimedia*, Ottawa, ON, Canada, Oct. 2001, vol. 9, pp. 623–625.
[25] R. Radhakrishnan, A. Divakaran, Z. Xiong, and I. Otsuka, "A content-adaptive analysis and representation framework for audio event discovery from "unscripte" multimedia," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–24, Jan. 2006.
[26] M. Irani, P. Anandan, and S. Hsu, "Mosaic based representations of video sequences and their applications," in *Proc. IEEE Int. Conf. Comput, Vis.*, Cambridge, MA, Jun. 1995, pp. 605–611.
[27] J. Wang and H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–638, Sep. 1994.
[28] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, Santa Barbara, CA, Jun. 1998, pp. 361–366.
[29] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.
[30] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graphics*, vol. 27, no. 3, pp. 1–9, Aug. 2008.
[31] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, Anchorage, AK, Jun. 2008, pp. 1–8.
[32] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking and Surveillance*, Beijing, China, Oct. 2005, pp. 65–72.
[33] Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. Circuits Syste. Video Technol.*, vol. 15, no. 10, pp. 1245–1256, Oct. 2005.
[34] I. Mani and M. Maybury, *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.
[35] K. Liu, E. Terzi, and T. Grandison, "ManyAspects: A system for highlighting diverse concepts in documents," *Proc. VLDB Endowment*, vol. 1, no. 2, pp. 1444–1447, Aug. 2008.
[36] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, Sep. 2008.
[37] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[38] S. Soatto, G. Doretto, and Y. Wu, "Dynamic textures," in *Proc. IEEE Int. Conf. Comput. Vis.*, Vancouver, BC, Canada, Jul. 2001, vol. 2, pp. 439–446.
[39] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, vol. 2, pp. 589–600.
[40] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, Feb. 2007.
[41] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[42] I. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. ACM SIGKDD*, Seattle, WA, Aug. 2004, pp. 551–556.
[43] Y. Wang, H. Jiang, M. Drew, Z. Li, and G. Mori, "Unsupervised discovery of action classes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, New York, Jun. 2006, pp. 1654–1661.
[44] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recogn.*, Cambridge, U.K., Aug. 2004, pp. 32–36.

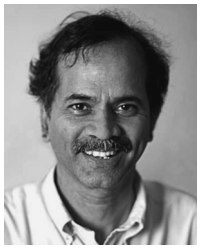
Nitesh Shroff (S'09) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 2007. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Maryland, College Park.

His research interests are in video processing, computer vision, and computational imaging.

Pavan Turaga (S'05–M'09) received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, India, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 2008 and 2009, respectively.

He is a Research Associate with the Center for Automation Research, University of Maryland, College Park. His research interests are in statistics and machine learning with applications to computer vision and pattern analysis. His published works include human activity analysis from videos, video summarization, dynamic scene analysis, and statistical inference on manifolds for these applications.

Dr. Turaga was the recipient of the Distinguished Dissertation Fellowship in 2009. He was selected to participate in the Emerging Leaders in Multimedia Workshop by IBM, New York, in 2008.



Rama Chellappa received the B.E. (Hons.) degree from the University of Madras, Madras, India, in 1975, the M.E. (Distinction) degree from Indian Institute of Science, Bangalore, India, in 1977, and the M.S.E.E. and Ph.D. Degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

Since 1991, he has been a Professor of Electrical Engineering and an affiliate Professor of Computer Science with the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director) and the Institute for Advanced Computer Studies (Permanent Member). In 2005, he was named a Minta Martin Professor of Engineering. Prior to joining the University of Maryland, he was an Assistant (1981–1986) and Associate Professor (1986–1991) and Director of the Signal and Image Processing Institute (1988–1990) with the University of Southern California (USC), Los Angeles. Over the last 28 years, he has authored or coauthored numerous book chapters and peer-reviewed journal and conference papers. He has coauthored and edited books on MRFs, face and gait recognition, and collected works on image processing and analysis. He has served as a co-editor-in-chief of *Graphical Models and Image Processing*. His current research interests are face and gait analysis, markerless motion capture, 3-D modeling from video, image and video-based recognition and exploitation, and hyper spectral processing.

Prof. Chellappa is is a Fellow of the International Association for Pattern Recognition and the Optical Society of America. He has served as the associate editor of four IEEE Transactions and as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He served as a member of the IEEE Signal Processing Society Board of Governors and as its Vice President of Awards and Membership. He has served as a General the Technical Program Chair for several IEEE international and national conferences and workshops. He is a Golden Core Member of the IEEE Computer Society and served a two-year term as a Distinguished Lecturer of the IEEE Signal Processing Society. He is serving a two-year term as the President of the IEEE Biometrics Council. He was the recipient of several awards, including an National Science Foundation Presidential Young Investigator Award, four IBM Faculty Development Awards, an Excellence in Teaching Award from the School of Engineering at USC, and two paper awards (coauthor of a Best Industry-related Paper and a Best Student Paper) from the International Association of Pattern Recognition. He received the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society. He also received a Technical Achievement Award and a Meritorious Service Award from the IEEE Computer Society. At the University of Maryland, he was elected as a Distinguished Faculty Research Fellow, as a Distinguished Scholar-Teacher, received the Outstanding Faculty Research Award from the College of Engineering. He also received an Outstanding Innovator Award from the Office of Technology Commercialization, University of Maryland, and an Outstanding GEMSTONE Mentor Award from the University of Maryland. He will be recognized as an outstanding ECE, Purdue University in September 2010.