

# A Spatiotemporal Decomposition Strategy for Personal Home Video Management

Haoran Yi<sup>a</sup>, Igor Kozintsev<sup>b</sup>, Marzia Polito<sup>b</sup>, Yi Wu<sup>b</sup>, Jean-Yves Bouguet<sup>b</sup>, Ara Nefian<sup>b</sup> and Carole Dulong<sup>b</sup>

<sup>a</sup>Computer Science Department, UIUC, Urbana, USA

<sup>b</sup>Intel Corporation, 2200 Mission College Blvd, Santa Clara, USA

## ABSTRACT

With the advent and proliferation of low cost and high performance digital video recorder devices, an increasing number of personal home video clips are recorded and stored by the consumers. Compared to image data, video data is larger in size and richer in multimedia content. Efficient access to video content is expected to be more challenging than image mining. Previously, we have developed a content-based image retrieval system and the benchmarking framework for personal images. In this paper, we extend our personal image retrieval system to include personal home video clips.

A possible initial solution to video mining is to represent video clips by a set of key frames extracted from them thus converting the problem into an image search one. Here we report that a careful selection of key frames may improve the retrieval accuracy. However, because video also has temporal dimension, its key frame representation is inherently limited. The use of temporal information can give us better representation for video content at semantic object and concept levels than image-only based representation.

In this paper we propose a bottom-up framework to combine interest point tracking, image segmentation and motion-shape factorization to decompose the video into spatiotemporal regions. We show an example application of activity concept detection using the trajectories extracted from the spatio-temporal regions. The proposed approach shows good potential for concise representation and indexing of objects and their motion in real-life consumer video.

**Keywords:** personal home video management, key frame, motion factorization, motion trajectory, spectral clustering.

## 1. INTRODUCTION

As the availability and cost of digital media capture and storage devices decrease, the size of personal media collections of pictures and videos increases dramatically. Media indexing and retrieval systems are very important to facilitate users to interact, browse and search those multimedia collections. However, most of existing media management systems currently rely on text annotation in the form of labels and tags or on other text that accompanies the media (e.g., HTML text related to images in the web or subtitle files associated with DVD content, etc.). Manual labeling of the media data is a very tedious and error prone process. It is therefore unlikely to expect this type of data to be present in consumer media. In the last decades, content-based image retrieval (CBIR) systems<sup>1-3</sup> emerged as an alternative approach where instead of relying on the keyword annotation for indexing and search, visual image features are directly used for search and retrieval. Up to now, this resulted in significant progress in image retrieval.

Compared to image data, however, video data is significantly larger in size and, most importantly, richer in multimedia content. Therefore, video indexing and retrieval is generally a more challenging task. Starting from 2001, the TREC video track (TRECVID)<sup>4</sup> brings together an international community of researchers in evaluation of content-based video indexing and retrieval. It has greatly facilitated the progress in content based digital video retrieval in several specific areas, primarily in news videos. There has also been significant amount of work in mining sports videos and several other domain specific applications. A great part of TRECVID's success is attributed to existence of common benchmarking infrastructure which greatly improves interaction between researchers.



**Figure 1.** Key frame selection example: the left figure shows middle frame of the video; the middle picture shows the key frame that has the feature vector closest to the average feature vectors of the video; the right picture shows the first frame of the video.

The situation with personal videos is strikingly different. This domain constitutes an important source of newly created video content. Development of solutions to help the management, search and retrieval of personal home videos is becoming a topic of great interest among researchers in universities and industry. At the same time this content is in many aspects different from videos used in TRECVID and similar efforts. The key differences are due to personal nature of media, mostly people oriented content, amateur video capturing techniques and little or no annotation.

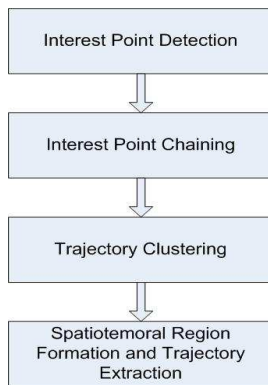
In<sup>5</sup> we developed a CBIR system and the benchmarking framework for image retrieval in personal image databases. In this work, we extend our image retrieval system to include personal home videos. The most direct way of extending the system is by extracting key frames for each video clip/shot and then using image retrieval techniques for search and retrieval. This method is simple and can leverage successful methods developed for image retrieval systems. However we can do better than only using a single frame for retrieval. The temporal dimension of the video provides valuable motion information for performing search, browsing and retrieval. Several approaches have been proposed to analyze the motion information for videos.<sup>6,7</sup> A coarse level measurement of the amount of motion content in the video is presented in.<sup>6</sup> This measurement is extracted from the motion vectors in MPEG encoded video. In,<sup>7</sup> a motion representation, *pixel change ratio map*, is proposed and the histogram from the *pixel change ratio map* is used as the feature to index the motion content.

In this paper, we present a new spatiotemporal decomposition strategy for personal home video content management. As in,<sup>8</sup> we seek a spatiotemporal decomposition of the videos. However, instead of clustering the whole cube of spatiotemporal pixels, we propose a *bottom-up* approach that combines interest point chaining, shape and motion factorization and image segmentation to decompose the video into a set of spatiotemporal regions (ST-regions). The decomposed ST-regions can be regarded as individual objects in the context of video indexing and retrieval. The trajectories extracted from the object ST-region demonstrate good potential for video retrieval of real-life consumer videos.

The remainder of this paper is organized as follows. In Section 2, we present the key frame extraction method. Section 3 details the *bottom-up* trajectory extraction and spatiotemporal decomposition approach. Experimental results are presented in Section 4. Finally, concluding remarks and future works are given in Section 5.

## 2. KEY FRAME EXTRACTION

Key frames provide abridged representation of the original video sequence. The video content indexing based on key frames is a well-known technique. The general approach is to first segment the video into shots with shot boundary detection algorithms. Then, key frames are extracted from the shots. A shot in a video sequence refers to a contiguous recording of a sequence of video frames depicting a continuous action in time and space. For personal home video clip collection, the video clips are taken after the camera is turned on and until it is turned off. Therefore, each individual clip represents a shot. But the frames in one clip have very wide variety in appearance due to the amateurish manipulation of the camera, e.g. jerky movement and shaking. This makes the key frame an important problem to solve.



**Figure 2.** Diagram for trajectory extraction and spatiotemporal decomposition.

Several heuristics have been used to select the key frame. For example, we can simply choose the first frame or middle frame as the key frames. We can also compute the average feature vector of all the frames and choose the frame whose feature vector is the closest to the average feature vectors. Figure 1 shows an example for key frame extraction with different heuristics. The left picture shows the key frame extracted as middle frame of the video; the middle picture shows the key frame that has the feature vector closest to the average feature vectors; the right picture shows the key frame as the first frame. The feature vector we use to compute the average feature vector is the 48 dimension color histogram, where the R, G and B components are quantized into 16 bins. From the visual inspective, the frame selected by average color histogram gives better representation of the scene content. Both the yard background and baby appeared to be better represented in this frame than in the other two. In section 4, we evaluate the performance of different key frame selection heuristics for video retrieval.

### 3. TRAJECTORY EXTRACTION AND SPATIO-TEMPORAL VIDEO DECOMPOSITION

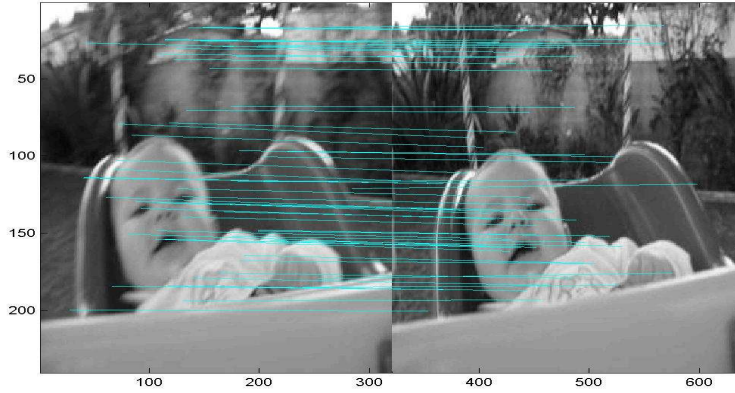
The motion features of a video sequence provide the easiest access to its temporal dimension and hence are of great importance for video content indexing. Trajectories are one of the most informative and important motion features. In fact, an object can be defined as a semantically coherent set of spatiotemporal regions containing trajectories describing motion relevant to the semantics of the object itself. Due to their importance, they have been standardized as a descriptor in MPEG-7.<sup>9</sup> However, the trajectory extraction is not an easy task for personal home videos due to complex scene structure, object shape, and irregular camera and object motion.

One direct approach for trajectory extraction is to use an explicit tracking module to track predefined objects through the video. But there are several shortcomings for this type of approaches: they require initialization of the object; it is very hard to deal with object disappearing from the scene and new object entering into the scene and occlusions are hard to deal with. Background subtraction can be used to avoid manual initialization of moving object. This approach has been widely used in surveillance videos,<sup>10</sup> where the cameras are stationary. However, the assumption of static camera for background subtraction approach is inappropriate for personal home videos where the camera is usually held by hand.

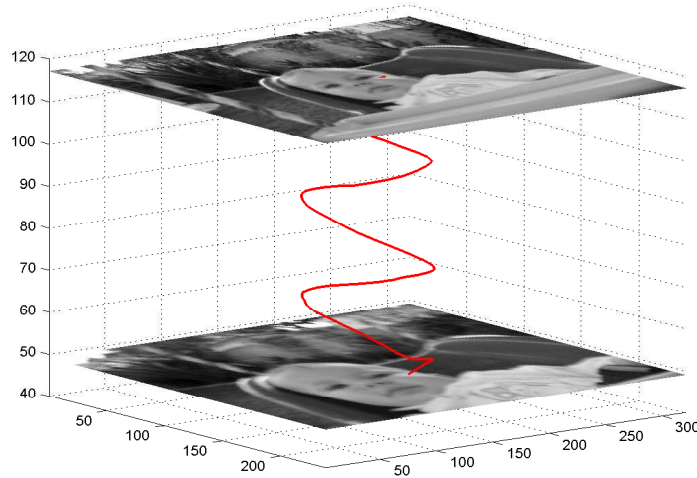
Because of these shortcomings, we proposed a *bottom-up* approach for trajectory extraction. The proposed approach consists of four stages: 1) interest points detection, 2) interest points chaining, 3) trajectory clustering and 4) spatiotemporal region formation and representative trajectory extraction. The diagram of the proposed approach is shown in Figure 2.

#### 3.1. Interest points detection

A wide variety of low-level interest points detectors exist in literature, such as Harris corner detector,<sup>11</sup> affine invariant point detector<sup>12</sup> and SIFT point detector.<sup>13</sup> Given the detected interest point, various local descriptors have been proposed such as high order image derivatives, generalized color moments, SIFT histograms, etc.



**Figure 3.** SIFT interest point detection and correspondence: left and right images shows the frame #100 and #103. The light blue lines link the corresponding points.



**Figure 4.** An example trajectories after SIFT point chaining.

The SIFT (*scale invariant feature transform*) point detector was originally introduced by Lowe.<sup>13</sup> Mikolajczyk and Schmid compare several interest point detectors in.<sup>14</sup> In their paper, the SIFT interest point detector and descriptor was demonstrated to be one of the most efficient and robust. Therefore, we choose SIFT detector as the interest point detector. Figure 3 shows the detected SIFT points on a baby swinging video. The left and right images show the SIFT points detected on frame #100 and frame #103. Most of the distinct points on the swinging baby (i.e. eyes, nose, mouth, etc.) and background (i.e. bushes, branches, etc) are detected.

### 3.2. Interest points chaining

After extracting interest points on every frame, the next step is “interest point chaining”. The aim is to link the individual interest point across adjacent frames to get a set of short trajectories.

To this extent we need to compute the similarity between interest points and construct the correspondence among them. The similarity measure between two interest points,  $i$  and  $j$ , is defined as the cosine of the angle between the two SIFT histogram descriptors.

$$\cos \theta_{i,j} = \mathbf{Des}(\mathbf{i}) \cdot \mathbf{Des}(\mathbf{j}) \quad (1)$$

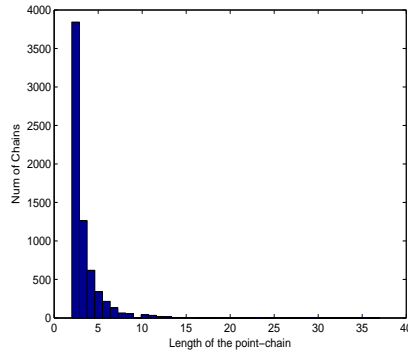


Figure 5. Length distribution histogram of the video trajectories.

where  $\mathbf{Des}(i)$  and  $\mathbf{Des}(j)$  are the normalized SIFT histogram descriptors for interest points  $i$  and  $j$ . The correspondence of the interest points is computed as follows. For each interest points,  $x$ , in frame  $n$ , we find the most similar point,  $y$  and the next most similar point,  $z$ , in the next frame  $n + 1$ . If the ratio,  $\frac{\cos \theta_{x,y}}{\cos \theta_{x,z}}$ , is larger than the threshold  $\alpha$ ,  $x$  and  $y$  are considered as a correspondence. The advantage of this approach is that there is no absolute threshold on the value of similarities and the extracted correspondences of points are highly distinctive, which makes the process very reliable. Once the frame-by-frame correspondence of the interest points is established, all the corresponding interest points are chained across frames to get a set of short trajectories. In practice, the interest point chaining can be done between frames with fixed interval rather than for every consecutive frames. For our implementation, the interest points are chained at 3 frame interval. Figure 3 shows the detected SIFT interest points and their correspondence on frame #100 and #103 for a “baby swing” video. The salient points on both the swinging baby (i.e. eyes, nose, mouse, etc.) and background (i.e. bushes, branches, etc) are detected and correctly matched.

Figure 4 shows the longest trajectory resulting from the point chaining on the example baby swing video. The interest points on the baby’s cheek are chained across different frames. The extracted trajectory is very reliable and lasts long enough (over 70 frames) to be used as a content descriptor. However, most the trajectories are short. Figure 5 shows the length distribution histogram of the video trajectories for the baby swing video. Most of the trajectories last less than 10 frames. These short trajectories are not good enough for motion content description. In the following subsection we describe how to apply trajectory clustering to group those short trajectories and achieve the spatiotemporal decomposition of the video. After trajectory clustering and spatiotemporal region formation, representative trajectories are selected from the salient object spatiotemporal regions. These trajectories give better description of the motion content than randomly selected trajectories.

### 3.3. Trajectory Clustering

There may be multiple motions within the videos and different trajectories belong to different motion groups, i.e. object and background etc. The goal of trajectory clustering is group the trajectories into independent motions such as object trajectories and background trajectories.

The measurement of how likely that two trajectories belong to the same motion group can be derived from shape matrix according to the multibody motion factorization framework described in.<sup>15</sup> In the following, we review the theory of 3D motion shape factorization.

Suppose there are  $N$  independently moving objects in the scene and each object contains  $n_i$  3D points. Their homogenous coordinates are represented as a  $4 \times n_i$  matrix  $\mathbf{S}_i$ ,

$$\mathbf{S}_i = \begin{bmatrix} x_i^1 & x_i^2 & \cdots & x_i^{n_i} \\ y_i^1 & y_i^2 & \cdots & y_i^{n_i} \\ z_i^1 & z_i^2 & \cdots & z_i^{n_i} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad (2)$$

When a linear projection (paraperspective, orthographic, affine etc.) is assumed, we can collect and stack the corresponding projected image coordinates  $(u, v)$  of these  $n_i$  points over  $F$  frames into a  $2F \times n_i$  matrix  $\mathbf{W}_i$ ,

$$\mathbf{W}_i = \mathbf{M}_i \cdot \mathbf{S}_i \quad (3)$$

where  $\mathbf{W}_i = \begin{bmatrix} u_{1,1} & \cdots & u_{1,n_i} \\ v_{1,1} & \cdots & v_{1,n_i} \\ u_{2,1} & \cdots & u_{2,n_i} \\ v_{2,1} & \cdots & v_{2,n_i} \\ \cdots & \cdots & \cdots \\ u_{F,1} & \cdots & u_{F,n_i} \\ v_{F,1} & \cdots & v_{F,n_i} \end{bmatrix}$  and  $\mathbf{M}_i = \begin{bmatrix} \mathbf{M}_{i,1} \\ \mathbf{M}_{i,2} \\ \cdots \\ \mathbf{M}_{i,F} \end{bmatrix}$  Each column of  $\mathbf{W}_i$  contains the observations for a

single point over  $F$  frames and each row contains the observed  $u$  coordinates or  $v$  coordinates for a single frame.  $\mathbf{M}_i$  is a  $2F \times 4$  matrix and  $\mathbf{M}_{i,f}$ , ( $f = 1, \dots, F$ ) is  $2 \times 4$  projection matrix related to object  $i$  in the  $f_{th}$  frame. Assume that each object is well formed, i.e. at least 4 non-coplanar points are detected on each objects (non degenerate case). Thus, the  $n_i$  columns of  $\mathbf{W}_i$  reside in a 4D subspace spanned by the columns of  $\mathbf{M}_i$ . If we put all the feature points from the different  $N$  object together into a  $2F \times P$  matrix  $\mathbf{W}$ ,

$$\mathbf{W} = [\mathbf{W}_1 \ \mathbf{W}_2 \ \cdots \ \mathbf{W}_N] = [\mathbf{M}_1 \ \mathbf{M}_2 \ \cdots \ \mathbf{M}_N] \cdot \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_N \end{bmatrix} \quad (4)$$

where  $P = \sum_{i=1}^N n_i$  is the total number of trajectories in the scene. Since the motions of all the  $N$  objects are independent, the rank of  $\mathbf{W}$  is  $4N$ . Given only the observation  $\mathbf{W}$ , we can determined  $\mathbf{M}$  and  $\mathbf{S}$  up to a linear transform at most, since  $\mathbf{M} \cdot \mathbf{A}$  and  $\mathbf{A}^{-1} \cdot \mathbf{S}$  would satisfy the factorization for any invertible  $\mathbf{A}$ . However, we can compute the shape matrix from  $W$  by singular value decomposition.

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (5)$$

where  $\mathbf{U}$  is  $2F \times 4N$ ,  $\mathbf{\Sigma}$  is  $4N \times 4N$ , and  $\mathbf{V}$  is  $P \times 4N$ . The shape matrix is computed as  $\mathbf{Q} = \mathbf{V}\mathbf{V}^T$ . In,<sup>15</sup> the authors prove

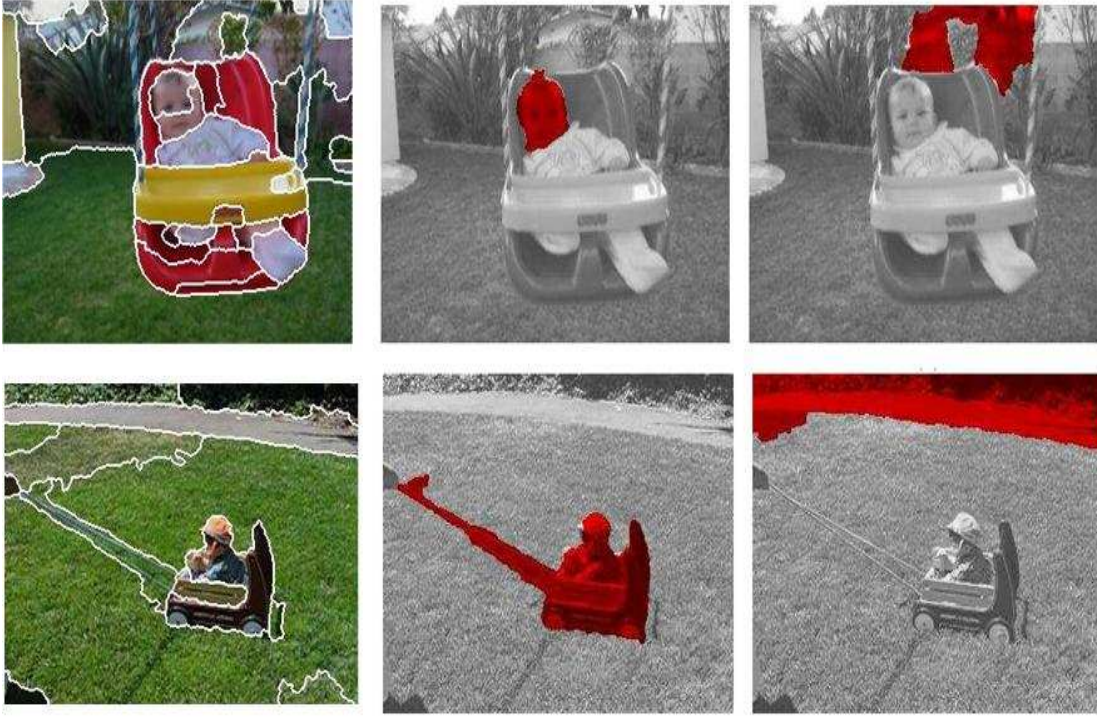
$$Q_{i,j} \begin{cases} = 0 & \text{if trajectory } i \text{ and } j \text{ belong to different object} \\ \neq 0 & \text{if trajectory } i \text{ and } j \text{ belong to same object} \end{cases} \quad (6)$$

The shape energy,  $Q_{i,j}^2$ , measures how likely two trajectories belong to the same motion group. The closer the value to 0, the less likely they are from the same motion. Given the  $Q^2$  matrix, it naturally fits into the graph based spectral clustering framework. The weighted trajectory graph is construct as follows: let the vertices represent the trajectories and assign the weight of edge  $e(i, j)$  to be  $Q_{i,j}^2$ . With this trajectory graph representation, normalized cut algorithm<sup>16</sup> can be applied to clustering the trajectories. As shown in<sup>16</sup> we need to solve the general eigen vector of the following equations,

$$\mathbf{L} \cdot \mathbf{q} = \lambda \mathbf{D} \cdot \mathbf{q} \quad (7)$$

where  $\mathbf{L}$  is the Laplacian matrix of the graph and  $\mathbf{D}$  is the diagonal matrix such  $D(i, i) = \sum_j W(i, j)$ . The second smallest generalized eigen vector gives the relaxed cluster membership value. By thresholding the relaxed cluster membership value, we get the clustering of the graph.

It is known in literature<sup>17</sup> that motion-shape factorization and the shape matrix,  $\mathbf{Q}$ , is not very robust. The quality of the result deteriorates dramatically when the noise level goes up. Therefore, the remaining problem is how to get a better measurement of the motion similarity than using the shape matrix alone. As shown in,<sup>17</sup> the trajectory clustering result can be improved by integrating other non-motion cues. The simplest and most useful cue is their geometric closeness. The observation is that the closer the two trajectories are, the more likely they belong to the same motion object. Besides the geometric closeness, other cues like visual similarity, segmentation can be considered too. But integration of those cues will increase the number of parameters to



**Figure 6.** Comparison of image-only segmentation (JSEG) and spatio-temporal segmentation on two example video sequences: the left column shows the image-only segmentation, the middle and right columns show the moving object and background regions using spatio-temporal decomposition.

estimate and different videos require different training set to get the appropriate parameters. In our current implementation, we only integrate the geometric closeness as a weighting factor of the motion similarity. The final motion similarity matrix is defined as follows:

$$W(i, j) = Q^2(i, j) \cdot g(\|T_i - T_j\|)$$

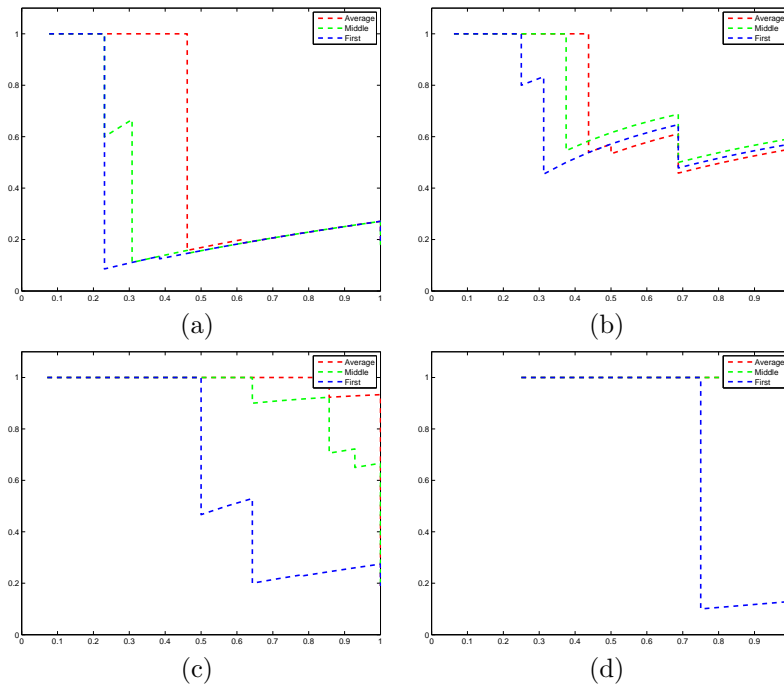
where  $g(\|T_i - T_j\|)$  is gaussian kernel weighting function. Plugging  $W$  in the N-cut equation (7), we can find the trajectory clusters.

After getting the trajectory clusters, the spatio-temporal regions are formed by grouping all image regions that trajectories in the cluster pass through. This process helps to alleviate the over-segmentation problem faced with many image segmentation algorithms. Figure 6 shows two examples of the extracted moving object and background region with the spatio-temporal region and those with image-only segmentation\*. The first column in Figure 6 shows the JSEG image-only segmentation result on these two videos. Both the main objects (i.e, baby, swing, kid, car etc.) and background regions (yard, bush, grassland etc.) are over-segmented. The middle and right columns show the extracted moving object and background regions by our trajectory clustering based STregion formation method. The moving “baby face”, “little car” and bush area in the background are extracted. The integration of motion cues help to enforce spatio-temporal coherence. As a result, the oversegmented regions are grouped into more complete moving object and background regions.

#### 4. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the proposed personal home video management system on real user videos. We collect 74 home video clips together with user’s annotation. The video collections are

\*JSEG<sup>18</sup> is used to obtained the image segmentation.



**Figure 7.** ROC curves for 4 query key phrase: (a) ‘swing’, (b) ‘activity gym’, (c) ‘back-yard’ and (d) ‘Halloween’.

taken over six months. The collection of video clips contains various events, e.g. baby swing, baby crawling, baby walking, birthday party, Halloween etc.

#### 4.1. Video retrieval by key frames

In this experiment, we compare three different video key frame extraction methods as described in Section 2. The first method is to choose the first frame as key frame. The second method is to choose the middle frame as key frame. The third method is to choose the frame whose color features is the closet to the average color feature of all the frames in the video. The color feature in use is a 48 dimension RGB histogram, where the R, G, and B are quantized into 16 bins.

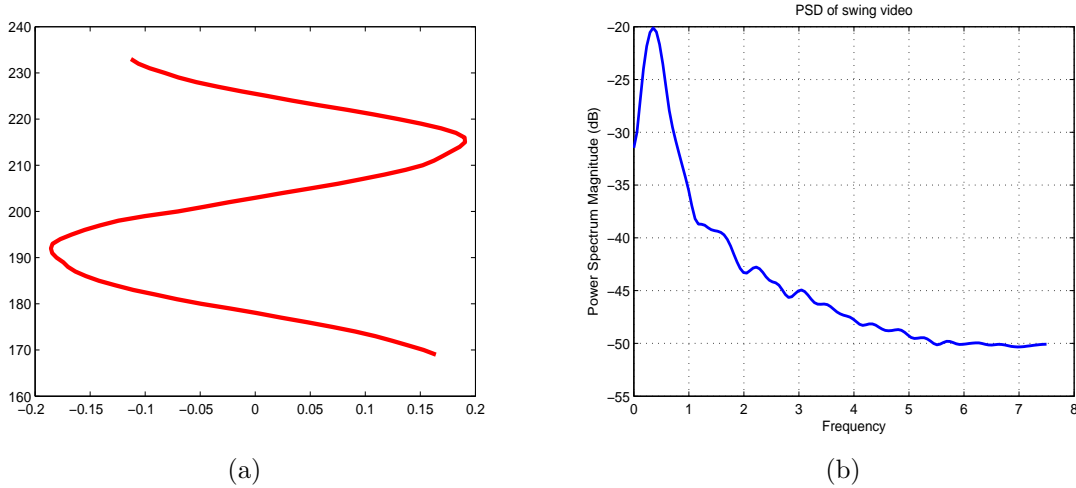
From the annotation file provided by the user, we found that there are four mostly used key words: ‘swing’, ‘activity gym’, ‘back-yard’ and ‘Halloween’. The key word ‘swing’ appears 13 times(17.6% of the total videos); ‘activity gym’ appears 16 times (21.6% of the total videos); ‘back-yard’ appears 14 times(18.9% of the total videos); ‘Halloween’ appears 4 times (5.4% of the total videos). We use these four key words as queries to do video retrieval. The ROC curves for the four queries are shown in Figure 7. We use the image retrieval benchmark framework developed in.<sup>5</sup> We simulate 2 rounds of user’s relevance feedback with 10 labeled samples (both positive and negative examples). SVM is used to learn from the relevance feedback and rerank the video clips. In order to achieve better performance, we use 128 dimension HSV color correlogram instead of the RGB histogram used in key frame selection. The color correlogram is more expensive to compute. But since we only calculated it on key frames, not for every frames, this computation is not much per video. The red, green and blue lines show the retrieval results with the average feature vector frame, middle frame and first frame respectively. The key frame extracted by average color histogram methods performs slightly better than the other two.

#### 4.2. Activity concept detection using trajectories

In this subsection, we present the initial experimental result of the bottom-up trajectory extraction and spatiotemporal video decomposition strategy described in Section 3. We illustrate the activity concept retrieval using spatiotemporal features extracted from trajectories. The concept chosen to retrieve is “baby swing”. The same data set described above are used for this experiment. 14 of 74 videos are “baby swing” videos and the



**Figure 8.** Key frames from of some example videos. The top row shows the “baby swing” videos. The bottom row shows other “non-swing” videos.

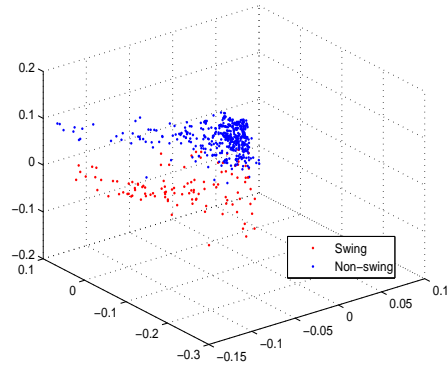


**Figure 9.** Example projected trajectory and its Power Spectrum Density (PSD): (a) the trajectory of the baby face spatiotemporal region in the most dominant moving direction.  $y$ -axis shows the frame number, and  $x$ -axis shows the magnitudes of movement. (b) The Power Spectrum Density (PSD) of the trajectory in (a).

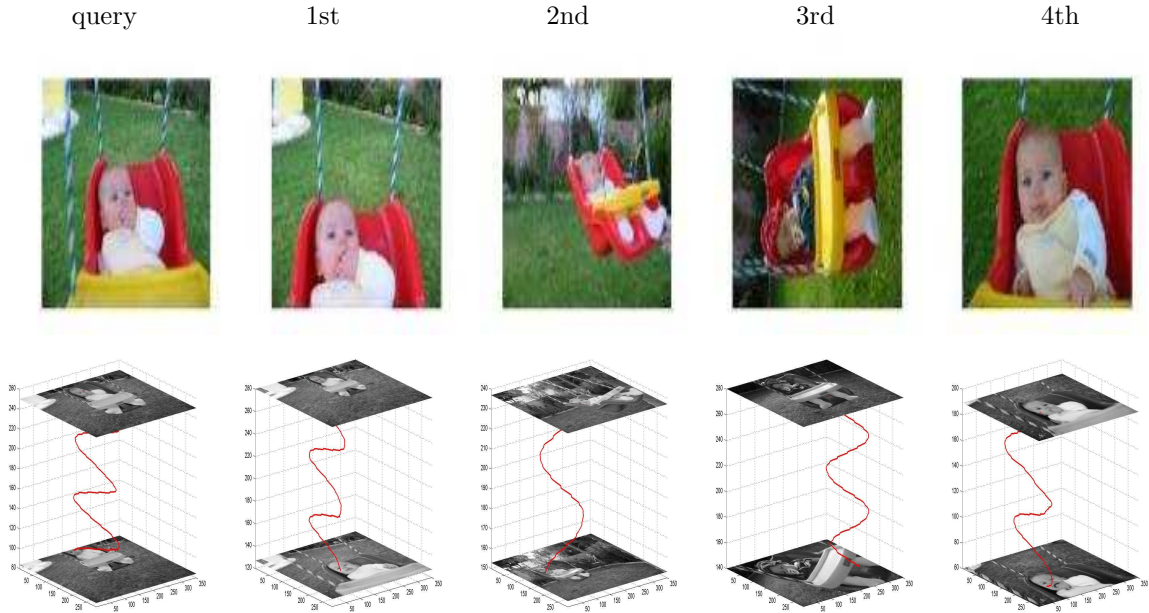
rest 60 are of various other activities such as “eating”, “babbling”, “singing” etc. Figure 8 shows several example key frames from the test video set.

Figure 9 gives an example illustration of the trajectory and its Power Spectrum Density (PSD). Figure 9(a) shows the PCA projected moving trajectories for the spatiotemporal baby face region shown in the middle figure on the top row of Figure 6: vertical axis is the time and horizontal axis is the principle direction of the trajectory in the  $x$ - $y$  plane. It is clear that the sinusoid shape trajectory gives a very good description of the swinging concept. Figure 4(b) shows the power spectrum density (PSD) of the swing trajectory. The dominant frequency is around 0.35 Hz which corresponds to around 3 second period. For ideal pendulum period  $T = 2\pi\sqrt{L/g}$ , the length of the string is about 2 meters, which roughly matches the actual physical length of the real swing.

As demonstrated in Figure 9, the PSD is a good feature for indexing of the “Swinging” concept. In the following, we present the details of using the PSD feature of the trajectories for “Swing” concept retrieval. The features for indexing are extracted in two steps: 1) the trajectories are projected onto the dominant moving direction using PCA and 2) the Power Spectrum Density (PSD) of the projected trajectories is extracted. The final feature vector consists of the magnitudes in the two orthogonal moving directions and the Power Spectrum Density vector (As we are using a 256 point FFT to compute the PSD, the final feature vector is of length 258.) The extracted feature index has desirable property such as rotation and shift invariant, which is important to



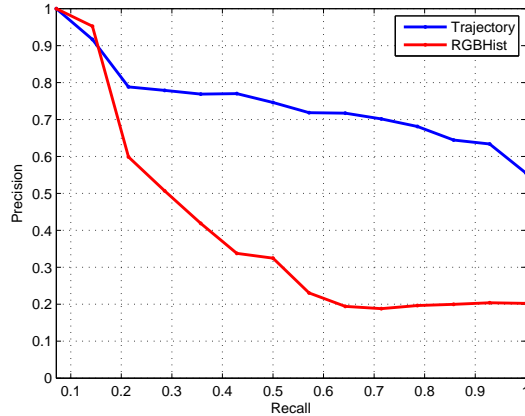
**Figure 10.** 3D plot of PCA projected trajectory feature vectors.



**Figure 11.** “Swinging” video retrieval result. The top row shows the key frames for one query example video and its top 4 retrieved videos. The bottom row shows the representative trajectories for each corresponding videos.

retrieve the swing video captured in different viewing angle.

For the video retrieval, we identify the moving object trajectory cluster by picking out the cluster that has the largest average moving magnitudes. Four longest trajectories from the identified moving object trajectory clusters are chosen as the representative trajectories for retrieval and their 258-D feature vectors are representative feature vectors. These 258-D feature vectors are very sparse as most of the middle and high frequency components in PSD are very close to zero. PCA is used to reduce the dimension of the feature vector. Figure 10 plots the PCA projected 3D feature vectors. The red (blues) dots denote the trajectories from the swing (non-swing) videos. As the trajectory clustering is not perfect, some non-swing trajectories from the background are clustered into the “swing” trajectories. However, even with these errors the swing trajectories (red dots) are well separated from the non-swing trajectories (blue dots) in Figure 10. During retrieval, the distance between two videos is calculated as the smallest distance from the set of trajectories from one video to the other. This method can tolerate the error from the miss-classified trajectories in the clustering stage.



**Figure 12.** Precision-recall curve for “Swinging” concept.

Figure 11 shows an example “swing” video retrieval result. The key frames for the query video and its top four retrieved videos are shown in the upper row. The corresponding representative trajectories for each videos are shown below the key frames. The “Swinging”s captured in different viewing angels and camera position are correctly retrieved. This demonstrate that our features are invariant to rotations and works well for the “swing” concept. Figure 12 compares the “Swinging” concept retrieval using trajectory feature with using color features. The reasons why we choose color feature for comparison are 1) its simplicity and wide usage in image/video retrieval and 2) the 14 swinging videos are taken when the baby was swinging in the same swing in the backyard, hence there are significant amount of color similarities. The color features are 48D RGB histogram calculated on key-frames. The blue curve in figure 12 shows the precision-recall curve using trajectory features. The precision of retrieval is 76% when the recall is 50%. The red curve in figure 12 shows the precision and recall curve for color features. The precision drops very quickly as the recall goes up. The precision drops to 21% when the recall is 50%. We observe that there are a few nearly duplicated videos among the swinging videos and the precision of the top retrieved result using 48D RGBHist is better than the trajectories. This indicates that color feature is good at detecting nearly duplicated video. However, it is not enough for detecting activity concept such as ‘swinging’. The trajectory feature does a much better job for swing concept detection. Although the experiment is limited, the spatio-temporal feature, i.e. trajectory, shows good potential for at least some motion specific activity concepts detection like “swing”.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we extend content-based image retrieval framework to include personal home videos. We present two different approaches for representation and retrieval of personal videos. The first one is based on key-frame extraction, while the second one leverages a spatiotemporal decomposition strategy. The key frame provides snapshot of the video clip; the spatiotemporal decomposition gives a more detailed description of the video content at object level and provides concise motion and temporal information representation, i.e., trajectories, for indexing. The proposed spatiotemporal decomposition is a bottom-up approach. It starts from SIFT interest point detection. The detected SIFT points are “chained” across frames to get a set of short trajectories. With the short trajectories, motion cue and geometric closeness are used to cluster the trajectories into independent moving groups. The clustering result gives the spatiotemporal region decomposition and spatiotemporal features, i.e. trajectories, for indexing. We demonstrate a potential usage of the spatiotemporal features for one example of activity (“swing”) concept retrieval by looking at its Power Spectrum Density (PSD). Future work will involve exploring different features extracted from the spatiotemporal decomposition and trajectories for video retrieval and more complex models for activity concept retrieval/detection.

## ACKNOWLEDGMENTS

This work was done when the first author was doing an internship at Intel Corporation Application Research Lab(ARL). We would like to thank all the members in ARL for their valuable comments and suggestion. We highly appreciate the video data and annotations provided by the ARL team members. We also thank the Intel Research Center in China (ICRC) for providing the video processing source code.

## REFERENCES

1. R. Veltkamp and M. Tanase, "Content-based image retrieval systems: a survey," Technical Report UU-CS-2000-34, Utrecht University, 2000.
2. H. Mueller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough, "Benchmarking image retrieval applications," *Proc. of the Seventh International Conference on Visual Information Systems, San Francisco, USA*, September 2004.
3. H. Mueller, W. Mueller, D. Squire, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," Technical Report 99.05, University of Geneva, 1999.
4. A. Hauptmann and M. Christel, "Successful approaches in the TREC video retrieval evaluations," in *Proc. of ACM Multimedia*, pp. 668–675, (New York, NY), October 2004.
5. J.-Y. Bouguet, C. Dulong, I. Kozintsev, and Y. Wu, "Requirements for benchmarking personal image retrieval systems," in *Proc. of the SPIE/IST Conference on Internet Imaging*, **7**, jan 2006.
6. K. Peker, A. Alatan, and N. Akansu, "Low-level motion activity features for semantic characterization of video," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2000.
7. H. Yi, D. Rajan, and L.-T. Chia, "A new motion histogram to index motion content in video segments," *Pattern Recognition Letters* **26**(9), pp. 1221–1231, 2005.
8. D. Dementhon and D. Doermann, "Video retrieval using spatio-temporal descriptors," in *Proc. of ACM Multimedia*, pp. 508 – 517, 2003.
9. B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG 7: Multimedia Content Description Language*, John Wiley & Sons, 2002.
10. A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. IEEE ICCV*, (Corfu, Greece), Sep 1999.
11. C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, (Manchester), 1988.
12. K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *European Conference on computer vision*, pp. 128–132, (Copenhagen), 2002.
13. D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision* **2**(60), pp. 91–110, 2004.
14. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *IEEE CVPR*, pp. 257–263, jun 2003.
15. J. Costeira and T. Kanade, "A multibody factorization method for independently moving-objects," *International Journal on Computer Vision* **29**, pp. 159–179, Sept 1998.
16. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8), pp. 888–905, 2000.
17. A. Gruber and Y. Weiss, "Incorporating non-motion cues into 3d motion segmentation," in *Proc. of ECCV*, (Graz, Austria), May 2006.
18. Y. Deng and B.S.Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, Aug 2001.