Interest Seam Image

Xiao Zhang*
Tsinghua University
andypassion.tech.officelive.com

Gang Hua
Nokia Research
ganghua@gmail.com

Lei Zhang, Heung-Yeung Shum
Microsoft Corporation
{leizhang, hshum}@microsoft.com

Abstract

We propose interest seam image, an efficient visual synopsis for video. To extract an interest seam image, a spatiotemporal energy map is constructed for the target video shot. Then an optimal seam which encompasses the highest energy is identified by an efficient dynamic programming algorithm. The optimal seam is used to extract a seam of pixels from each video frame to form one column of an image, based on which an interest seam image is finally composited. The interest seam image is efficient both in terms of computation and memory cost. Therefore it is able to power a wide variety of web-scale video content analysis applications, such as near duplicate video clip search, video genre recognition and classification, as well as video clustering, etc.. The representation capacity of the proposed interest seam image is demonstrated in a large scale video retrieval task. Its advantages are clearly exhibited when compared with previous works, as reported in our experiments.

1. Introduction

The far reach of the internet has created gigantic amount of online videos. This is largely due to the ever more popular social video sharing activities supported by online social media web-sites, such as YouTube. Automatic content analysis of this enormous amount of video data becomes an emerging need, which may greatly facilitate users to manage, share, visualize, and search the relevant visual content.

Nevertheless, the magnitude of Internet video data imposes tremendous amount of scalability challenges, both in terms of computation and memory usage. To address them, an efficient yet effective visual representation of the video is indispensable. The visual representation must be computationally *super efficient* to be able to scale up to web-scale data. Moreover, the visual representation should be informative to differentiate between different video content. Last but not least, the visual representation should be compact to keep memory cost in an affordable manner.

*The main work of this paper is performed when Xiao Zhang is a research intern at Microsoft Bing search mentored by Gang Hua.

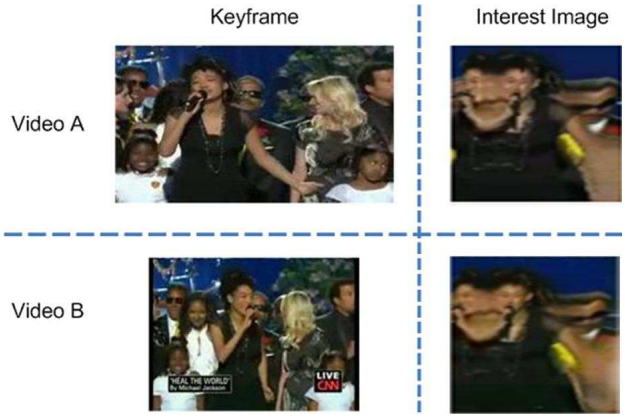


Figure 1. Key-frames and interest seam images from a pair of near duplicate videos. The key-frames contain same object with different poses which are difficult to match. In contrast, interest seam images are near duplicates which could be matched easily.

Visual representation from previous approaches often can not satisfy all the three conditions, and therefore are incapable of dealing with video corpus in the Internet scale. Some approaches [11, 8] extract spatiotemporal local invariant features from the raw video, from which a visual representation, e.g., a *bag-of-visual-feature* or a more complicated statistical model, is induced. Although the feature extraction may be sped up by using computational paradigm such as integral videos [6], still they are not efficient enough to handle billions of videos.

Some other approaches convert video content into fingerprints for fast processing. Most of them, such as the color signature and video histogram [9], are based on global histogram of visual features. They usually discard the rich temporal context in a video sequence, neither are they robust to visual transformations such as lighting enhancement, contrast sharpening, and gamma rectification, etc..

The third category of approaches firstly generate a condensed synopsis of the video, from which low level features are extracted to build the final visual representation for the video. The video synopsis could be generated, for example, from key-frames of a video shot [7], or from temporal

slice [12], which is a set of two dimensional images extracted along the time dimension of a video sequence. Since low-level features are only extracted from the synopsis of the video [2, 17, 16], approaches in the third category have great potential to be both computationally and memory efficient, and therefore to be scalable to web-scale data.

However, how to ensure the stability and repeatability of key-frame detection is an open problem. This is exemplified in the first column of Fig. 1, where key-frames extracted from near duplicate videos are quite different. Besides, key-frame based synopsis neglects temporal information, which is obviously very valuable for video content analysis tasks. Meanwhile, temporal slice, as well as space-time scene manifolds [20], may also discard valuable spatiotemporal visual information since no saliency information is accounted for.

We propose *interest seam image* for scalable video content analysis. It generates a compact synopsis of a video shot and keeps both important spatial and temporal information in a computationally efficient manner. A spatiotemporal energy map is constructed firstly. Then an optimal vertical seam which encompasses the highest energy is identified and used to extract a seam of pixels from each video frame. These seams of pixels are placed column by column to compound the interest seam image. In the second column of Fig. 1, we present interest seam images extracted from the videos. It is shown that the interest seam images are visually very similar to each other, which is in contrast to the key-frames extracted.

The proposed interest seam synopsis is highly informative and efficient, which can function as the basis for a large variety of applications such as video search, video genre recognition, classification, and clustering, etc.. We demonstrate its advantages in an application of large scale near duplicate web video retrieval. Our video retrieval system extracts local invariant features from each interest seam image. A MinHash scheme [2] is further leveraged to build efficient inverted file index [14, 4] for the video database. The inverted list for each MinHash key is sorted based on the temporal context of interest seam images, and hence achieves very efficient retrieval performance.

We further propose a novel post verification scheme, namely scene verification, to improve the retrieval accuracy. In contrast to geometric verification in image recognition, which assumes the visual objects are undergone a rigid transformation, scene verification is performed in a holistic level by leveraging global descriptors such as GIST [13, 3] with no additional assumption.

Therefore, the contributions of this paper are three folds: 1). We propose *interest seam image*, a novel video synopsis method. 2) Based on interest seam image, we employ MinHash and design a scheme for efficient inverted file indexing of the video corpus by taking the temporal context into consideration. 3). We propose *scene verification*, a novel post-verification method to further improve the retrieval accuracy, speed, memory consumption, and scalability.

2. Interest seam image

A video clip could be considered as a 3 dimensional spatiotemporal cube. Interest seam image is a synopsis of the clip which represents the most informative 2 dimensional curvature manifold in this cube. A naive method, namely temporal slice [12], obtains a slice by getting one column per frame and concatenate them together as an image, which might be uninformative and could not represent the video content very well, as shown in the bottom right of Fig. 2. The reason resides in the fact that the spatiotemporal informativeness of the pixels are not taken into consideration. To improve this method, we need to define an energy function in the image plane to measure the spatiotemporal saliency of each pixel, and then pick a group of pixels with the highest energy. Since our goal is to identify an efficient and compact visual synopsis of the video content, we propose to define the group of pixels to be vertical seams on consecutive frames which encompasses the highest energy. This leads to the proposed *interest seam image*, which provides flexible and rich representation, and is yet efficient to compute.

2.1. Seam based video synopsis

To obtain an interest seam image, for each frame I, we extract a "seam" $\mathbf{s}$ and concatenate seams of pixels from all frames column by column to form the interest seam image. Formally, let the resolution of input video be $m \times n$, where m is the width and n is the height of I, following the definition of [1], a vertical *seam* is defined as

$$\begin{aligned} \mathbf{s} &= \{s_i\}_{i=1}^{n} = (x(i), i)_{i=1}^{n} \\ & s.t. \quad \forall i, \ |x(i) - x(i-1)| \leq 1 \end{aligned} \tag{1}$$

where x is a mapping $x : [1, ..., n] \rightarrow [1, ..., m]$. That is, a vertical seam $\mathbf{s}$ is an 8-connected path of pixels in the image from top to bottom, containing only one pixel in each row of frame I. The pixels on the path of the seam $\{s_i\}_{i=1}^{n}$ will therefore be $\mathbf{I_s} = \{\mathbf{I}(s_i)\}_{i=1}^{n} = \{\mathbf{I}(x(i), i)\}_{i=1}^{n}$

With this definition, we proceed to extract interest seam image from a sequence of video frames. Simply obtaining a seam with the highest energy from each frame separately will introduce serious discontinuity into the resulted interest seam image, which may render it very difficult to extract meaningful low level features. Therefore, we choose to extract a seam in the image plane that is of the highest energy aggregating the saliency information from all the video frames. That is, the configuration of the seams in different video frames are the same. We proceed to introduce the aggregated spatiotemporal energy function we adopted.

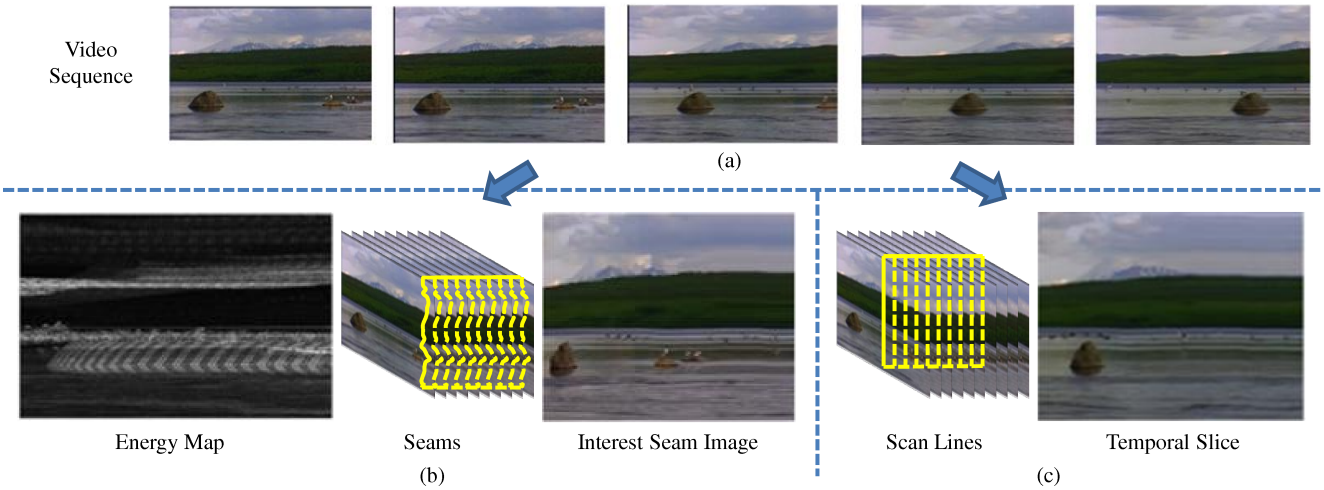


Figure 2. Interest seam image (bottom left) v.s. temporal slice(bottom right): interest seam image incorporates more information about the original video than temporal slice(best viewing in color format). Here the energy map visualizes E_{ST} in Eq.3

2.2. Spatiotemporal energy map

The visual saliency energy function we adopted is

$$E(i,j) = E_{ST}(i,j) + \beta E_{prior}(i,j), \tag{2}$$

where $i \in [1,n]$, and $j \in [1,m]$. Recall from the beginning of Sec. 2.1, m and n are defined as the width and height of the video frames. The first component of the energy function, E_{ST}, takes both spatial and temporal saliency information into consideration. This is done by computing the gradient and motion information in each frame independently and then use an operator f (e.g. average, max, median, etc.) to aggregate energy value at each pixel location from different frames, which is a generalization of the formulation of [15]. Specifically, given a video sequence $\{I_t\}_{t=1}^T$, its spatiotemporal energy function E_{ST} is defined as follows

$$E_{ST}(i,j) = (1-\alpha)E_S(i,j) + \alpha E_T(i,j) \tag{3}$$

$$E_S(i,j) = f(\{|\frac{\partial}{\partial x}I_t(i,j)| + |\frac{\partial}{\partial y}I_t(i,j)|\}_{t=1}^T) \tag{4}$$

$$E_T(i,j) = f(\{|\frac{\partial}{\partial t}I_t(i,j)|\}_{t=1}^T) \tag{5}$$

where $\alpha \in [0,1]$ balances the spatial and temporal energies.

Since we prefer seams with high energy, solely relying on E_{ST} may be prone to noises. Therefore, we introduce E_{prior}, a spatially central Gaussian prior, to the energy function in Eq. 2. It is defined as

$$E_{prior}(i,j) = exp(-(\frac{j-\frac{m}{2}}{\sigma m})^2) \tag{6}$$

The introduction of this prior is based on our observation that most web videos will place the most important and interesting content in the central part of video frames. Indeed, this is probably true for almost any video framing.

Therefore it is reasonable for the seam to bias toward the central part of the video frame, which in general enhance the repeatability of the seam, as demonstrated in our experiments. The strength of the Gaussian central prior is adjusted by $\beta \geq 0$ in Eq. 2. The effectiveness of Gaussian prior is tested in Sec. 5.1 and the sensitivity of the interest seam image against β is tested in Sec. 5.4.

2.3. Interest seam image idenfication

Given E, the energy that a seam contains is

$$E(\mathbf{s}) = \sum_{i=1}^{n} E(s_i) = \sum_{i=1}^{n} E(x(i), i). \tag{7}$$

We look for an optimal seam with highest energy, i.e.,

$$\mathbf{s}^\star = \max_{\mathbf{s}}(E(\mathbf{s})) \tag{8}$$

Let $\mathbf{s}_k = \{s_i\}_{i=1}^k$ be a sub-seam of $\mathbf{s}$, we denote

$$E_m(p,k) = \max_{\mathbf{s}_k : x(k)=p} \sum_{i=1}^{k} E(x(i), i). \tag{9}$$

as the maximum energy up to seam position (p,k) in $\mathbf{s}_k$, it is easy to figure out that

$$E_m(p,k+1) = E(p,k+1) + \max_{\Delta \in \{-1,0,1\}} \{E_m(p+\Delta, k)\} \tag{10}$$

From Eq. 10, the optimal seam $s^\star$ can be obtained efficiently using dynamic programming.

With the optimal seam $\mathbf{s}^\star$, the interest seam image is obtained by placing seams of pixels from consecutive frames

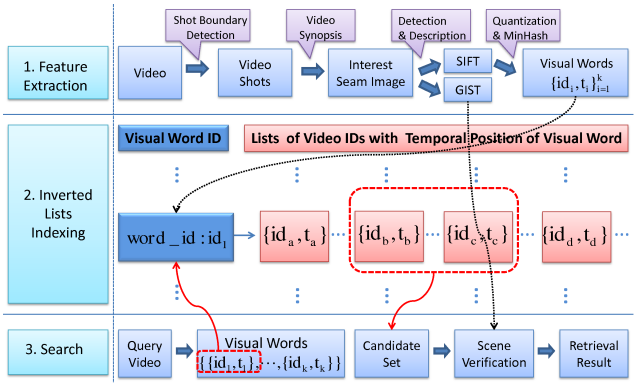


Figure 3. The retrieval system pipeline: as each inverted list is sorted by the temporal positions where the visual words occur in the interest seam images of the videos(i.e. $t_a < t_b < t_c < t_d$), only a small subset(i.e. the middle part) of the list need to be visited, because $t_a < t_1 - c < t_b < t_c < t_1 + c < t_d$.

column by column. Examples of energy map, seam and interest seam image are shown in Fig. 2. Comparing with temporal slice in this figure, we could see that interest seam image contains more information about the original video, such as the two mountain peaks in the top and the bird standing on the stone in the middle of the lake.

3. Application to video retrieval

Although interest seam image could be potentially applied in a variety of large scale video content analysis tasks, we focus on large scale near duplicate web video shot retrieval to demonstrate its advantages. The goal is to find similar web video shots which has undergone certain modifications such as logo/caption insertion, color change, aspect ratio change, and encoding format change etc. Some examples of near duplicate videos are shown in Fig. 5 in the experiment section.

Our video retrieval pipeline is shown in Fig. 3. In offline stage, the system automatically segments the database videos into shots [19]. Next, interest seam image is computed for each shot, from which both local feature(MSER [5]) and global feature(GIST [13]) are extracted. SIFT [10] is used to describe the gradient information in the neighborhood of the local features. To make the feature more compact, we quantize SIFT descriptors using vector quantization [14] to convert an interest seam image to a *bag of visual words* representation, i.e.,

$$V = \{P_i^V\}_{i=1}^N, P_i^V = \{id_i^V\}, \tag{11}$$

where each local feature P_i^V is represented by a visual word ID (i.e. id_i^V) obtained during the quantization process. The problem of this representation is that N is often very large and varies for different interest seam images. Therefore, using k hash functions $\{H_j\}_{j=1}^k$, we convert an interest seam image into a k dimensional *MinHash signature*

$$V = \{P_{m_j}^V\}_{j=1}^k, P_{m_j}^V = H_j(\{P_i^V\}_{i=1}^N), m_j \in \{1, ..., N\} \tag{12}$$

where the j^{th} dimension is a selected visual word obtained by applying the j^{th} hash function on the entire bag of visual words. The design of $\{H_j\}_{j=1}^k$ and the theoretical properties of MinHash could be found in [2]. Note that the dimensionality of MinHash signature, i.e. k, determines the memory cost of indexing. Therefore, to reduce memory cost, k is usually set to be much smaller than N in Eq. 11.

After the MinHash algorithm is applied, an indexing structure is built to quickly return a small candidate set of relevant videos for a query. Each inverted list is sorted based on the temporal context of the interest seam image for efficient access. Finally, a scene verification component using GIST features extracted from interest seam images is used to re-rank this candidate set. The details will be introduced in following subsections.

3.1. Inverted index with temporal context

In this section, we introduce the technical detail of the efficient inverted file indexing scheme with temporal context, which effectively improves retrieval accuracy and speed by leveraging the temporal information contained in interest seam image.

Recall from previous sections that the k^{th} column in an interest seam image is made up of a "seam" generated from the k^{th} frame in the video. Therefore the X axis of interest seam image could be considered as a *temporal axis*. Denote the position of local feature $P_{m_j}^V$ on temporal axis to be $t_{m_j}^V$, then the MinHash signature for interest seam image in Eq. 12 is expanded as

$$V = \{\{id_{m_j}^V, t_{m_j}^V\}\}_{j=1}^k. \tag{13}$$

The similarity between the query video and a database video is measured as

$$Sim(Q, V) = \sum_{j=1}^{k} I(id_{m_j}^Q = id_{m_j}^V, |t_{m_j}^Q - t_{m_j}^V| < c), \tag{14}$$

where $I(\cdot)$ equals 1 if all conditions in the parenthesis holds and 0 otherwise, and the parameter c introduces flexibility to this similarity measure to a certain degree. It confines that only local features with similar temporal positions could be matched in our system.

The temporal constrained similarity in Eq. 14 shares similar idea with spatial verification for large scale image search [14]. Both methods utilize spatial configuration of local features to improve accuracy. The difference is, the positions of local feature on X and Y axis are equally important for spatial verification, while for interest seam image,

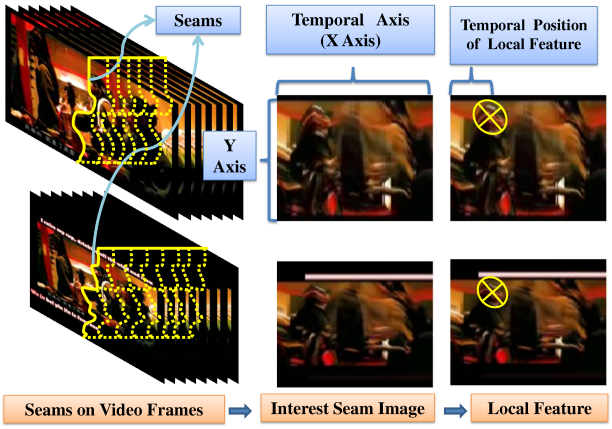


Figure 4. Illustration of the temporal axis in interest seam image. Because of the letterbox effect (the black bars on the top and bottom borders of frame) in the two videos, the same visual words (yellow eclipses) are located in different positions on Y axis. However their temporal positions are almost the same.

the position of local feature on temporal axis is much more stable to verify than its position on Y axis, as illustrated in Fig. 4. This is because letterbox effect, cropping and other editing in image plane is common so local features in near duplicate interest seam images are highly possible to be located on different positions on Y axis.

Besides the increase of accuracy, the efficiency of the video indexing structure is also improved. In the proposed indexing structure, the MinHash signatures of database videos are organized by inverted lists, as shown in the middle of Fig. 3. Inverted list index has an entry for each visual word in the database followed by a list of all the videos (and the temporal position in the interest seam image of each video) in which the visual word occurs. The advantage of the inverted lists is that a linear scan for all the database is avoided, i.e. only lists corresponding to the visual words of the query need to be scanned. However, for a large scale database, the lists themselves are very long and a linear scan over the entire list is slow itself.

In the proposed video indexing algorithm, temporal information is leveraged to sort the entries in each list thus by binary search we could efficiently identify a small subset of the list, in which the temporal positions of the visual words are similar to that of the query video. For example, in Fig. 3, as the first dimension of the MinHash signature of query video has value $\{id_1, t_1\}$, thus the list corresponding to visual word id_1 is looked up. According to Eq. 14, we only want to return videos containing visual word id_1 with temporal position t subject to $(t_1 - c) < t < (t_1 + c)$. Therefore instead of enumerating all the entries in the list, only the middle part of that list need to be scanned. In the experiments, the retrieval speed of this indexing structure is orders of magnitude faster.

3.2. Scene verification

After a small candidate set of videos is generated by local feature based index, a re-ranking mechanism using scene descriptor is added to further improve the precision of top search results. This is achieved by extracting GIST descriptors G from interest seam images [13, 3]. Then we compute the Cosine similarity of the scene descriptors between the query video and each video in the candidate set, i.e.,

$$Sim_{Gist}(Q, V) = \frac{G_Q \cdot G_V}{\|G_Q\|\|G_V\|}. \quad (15)$$

We then use a linear combination of the two similarities defined in Eq. 15 and Eq. 14 as the final ranking score to rerank the videos in the candidate set, i.e.

$$R(V) = \gamma Sim_{Gist}(Q, V) + (1 - \gamma) Sim(Q, V) \quad (16)$$

The higher $R(V)$ is, the more relevant V is to Q. This design is based on our observation that the visual scenes of the interest seam images extracted from near duplicate videos are usually very similar, although there may be subtle local differences. Therefore, performing scene verification could improve the search accuracy.

A nice property of this approach is that, with the help of the scene verification, empirically we find that not only the retrieval accuracy is improved, but the dimensionality of MinHash signature (i.e. k in Eq. 12) could be greatly reduced as well, which implies that much less information nodes need to be stored and visited in the inverted lists. Therefore, the memory efficiency, retrieval speed and accuracy are improved simultaneously.

4. Dataset and evaluation criterion

To conduct large scale video retrieval experiments, 10382 videos, which covers a variety of genres including TV shows, soap operas, movie trailers and MTVs etc., are crawled from Internet and automatically segmented [19]. Among them, two labeled dataset **Qr** and **Qt** are used as queries, with 400 video shots and 1500 video shots respectively, and a dataset **D** with 246551 video shots are used as distracters to further test the scalability of our algorithm. Each shot in our collection ranges from 2 to 10 seconds.[1]

Qr: The near duplicate videos in this dataset are videos that contain same content but have undergone a combination of transformations such as scaling, compression, color change, caption overlay, and logo insertion, etc., as shown in Fig. 5. We manually label the near duplicate relationships and use

[1]Please contact the authors for downloading the query and distractor(kindly provided by Dr. Linjun Yang) dataset.

Figure 5. Examples of near duplicate videos in **Qr**

these ground-truth label to compare the performance of our system with state-of-the-art systems in real world settings.

Qt: A group of 150 video shots is randomly selected from the crawled videos. Each of them undergoes several artificial transformations including scale change, horizontal mirror, blur, contrast stretch, and gama correction, which results in 9 transformed versions of each original video. Therefore, the dataset contains 1500 videos in total. In our experiments, the original 150 video shots are used as queries to test whether the transformed ones could be successfully retrieved. This dataset is used to compare the interest seam image synopsis with key-frame based methods.

With these query datasets, two groups of experiments are conducted, in which the evaluation strategy is similar to the leave-one-out strategy commonly used in object recognition. For the first group, each video shot from **Qr** is used in turn to query the database which combines **D** and **Qr**(the query video is excluded). The other group of experiments is conducted similarly but **Qr** is replaced by **Qt**. We compute Average Precision by averaging precision over all recall levels for each query. Then we get the mean value of average precisions for all queries, namely Mean Average Precision(**mAP**), to serve as the main performance evaluation criterion, which is also widely adopted for multimedia retrieval evaluation [18, 14].

5. Experiments

In this section, we present a variety of experiments conducted to demonstrate the effectiveness and efficiency of the proposed video representation. Firstly, a qualitative experiment is carried out to demonstrate the stability of the "seam" against various video transformations in Sec. 5.1. Then, the performance of the proposed video synopsis and retrieval system is compared with several state-of-the-art methods in Sec. 5.2. After that, the parameter sensitivity of the proposed algorithms are tested in Sec. 5.3 and Sec. 5.4. Finally, the strength of scene verification and inverted index with temporal context are investigated in Sec. 5.5 and Sec. 5.6.

Figure 6. Illustration of the stability of the seams (yellow curves) against video transformations. Videos in the second row are transformed from the first row.

Video Representation	mAP	
	MSER	GIST
Interest Seam Image	**0.9699**	**0.9766**
Keyframe	0.894	0.9062

Table 1. Performance comparison between interest seam image and key-frame on local feature(MSER) and global feature(GIST). The evaluation is carried on transformed queries(**Qt**).

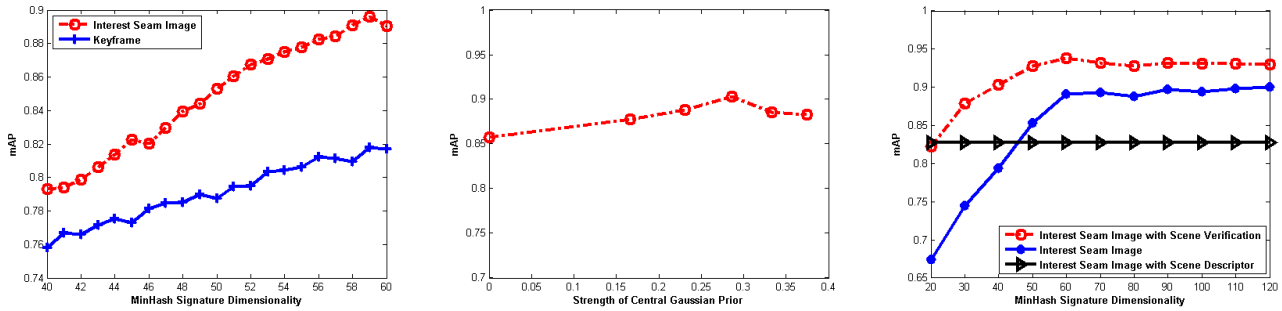
Before introducing the experimental results, we briefly discuss the implementation details of the proposed algorithms. All experiments are carried out on a 2.33GHZ PC. The β, α, σ, c, γ, in Eq. 2, Eq. 3, Eq. 6, Eq. 13, Eq. 16, are set to be 40, 0.7, $\frac{\sqrt{2}}{7}$, 30, 0.7, respectively. The dimensionality of MinHash signatures is 60. Besides, in all experiments, *max* is used to implement operator f in Eq. 4 and Eq. 5, and each dimension of GIST is quantized to an integer to reduce memory cost.

5.1. Stability of seam

In this experiment, we test the stability of the seam against different kinds of video transformations, including aspect ratio change, contrast stretch, gamma correction, and a more complicated transformation combining contrast stretch and logo insertion. The result is visualized in Fig. 6. From this figure, we could see the positions of the seams are quite stable thus the interest seam image composited from seams is suitable for near duplicate video matching.

5.2. Retrieval accuracy

To demonstrate the superiority of interest seam image over key-frame, we compare them with different low level features, using **Qt** as testing data. The result is shown in Table. 1. It is clear that no matter we use MSER [5] or GIST [13], the accuracy of interest seam image based method is much higher than the key-frame based



(a) (b) (c)

Figure 7. (a). Performance comparison between interest seam image and key-frame with changes in the MinHash signature dimensionality; (b) retrieval accuracy of interest seam image with changes in the strength (i.e. β in Eq. 2) of central Gaussian prior; (c). performance comparison among interest seam image, interest seam image with scene descriptor, and interest seam image with scene verification

Approaches	mAP	Global	Local
Interest Seam Image + Scene Verification	**0.9401**	Y	Y
Interest Seam Image	**0.896**	N	Y
Keyframe + MinHash [2]	0.8179	N	Y
Keyframe + GIST [3]	0.8472	Y	N
Temporal Slice [12]	0.8388	Y	N
Hierachical Framework [18]	0.8828	Y	Y

Table 2. Performance comparison between proposed algorithms and state-of-the-art methods on **Qr**. The last two columns indicate whether each method uses global/local feature. And "Interest Seam Image" means the proposed retrieval system without scene verification.

approaches.

We also compare the proposed video retrieval system with several recent methods [2, 18, 12, 3] on **Qr**. In [2], interest points were extracted from key-frames and indexed by Min-Hash. In [18] local and global features extracted from key-frames were combined in a hierachical framework to improve efficiency and preserve accuracy. GIST was used in [3] as global feature to index web images and good result is achieved. Therefore, we also used key-frame combined with GIST as a baseline in the experiment. Finally, we compared interest seam image with [12] which used temporal slice. The comparison results are summarized in Table. 2.

From the evaluation result, it could be seen that only with local feature, interest seam image already achieve 0.896 retrieval accuracy, which is better than all competing methods, including [18] which utilizes both global feature and local feature. After the scene verification based reranking step is added, the accuracy of the proposed video retrieval system is significantly further improved to reach 0.94.

5.3. Impact of dimension of MinHash signature

Recall from Sec. 3.1 that the signature dimensionality k in Eq. 12 determines both the representation capability of the signature and the memory cost of the retrieval system. Therefore we vary k, and compare the accuracy of the proposed retrieval system with key-frame combined with MinHash [2] on **Qr**. The result is shown in Fig. 7(a). This figure shows that interest seam image based indexing performs better than key-frame on all levels of k. Besides, it also shows that even with a small dimensionality (e.g. 60), our retrieval system could achieve good retrieval result, i.e. 0.896 in terms of mAP. This indicates that our video retrieval system could be scaled to web video dataset.

5.4. Impact of Gaussian central prior

The impact of the parameter β in Eq. 2 is tested on **Qr**. The result is shown in Fig. 7(b). We could see that if no Gaussian prior is enforced on the generation of seam, the seam will become less repeatable and lead to performance degradation. Increasing the strength of Gaussian prior will make an remarkable improvement on the retrieval accuracy(mAP improved from 0.8577 to 0.9029) but if the Gaussian prior is too strict, the seam will be less flexible and hurt the retrieval results.

5.5. Impact of scene verification

To fully justify the benefit of scene verification, we compare the proposed retrieval system with two of its variants: interest seam image without scene verification(the curve for "Interest Seam Image" in Fig. 7(c)) and interest seam image with only scene descriptor. The first variant only uses local feature, the second variant only uses scene descriptor, while the proposed retrieval system combines both features. The comparison result is shown in Fig. 7(c). We could see that the scene verification approach performs much better than using either kind of feature individually. Also, it could be seen that with the help of scene verification, the accuracy of the proposed retrieval system is improved even with much less dimensions of MinHash signature, which implies that

Approaches	mAP	Memory	Speed
MinHash	0.8886	720 Bytes	4.09s
MinHash + Temporal Context	0.9165	1080 Bytes	0.47s
MinHash + Temporal Context +Scene Verification	**0.9401**	**480 Bytes**	**0.24s**

Table 3. The mAP and cost comparison of MinHash, MinHash with temporal context, MinHash with temporal context and scene verification.

both accuracy and memory efficiency of the retrieval system are improved.

5.6. Efficiency comparison

In this section, we compare the efficiency of the proposed retrieval algorithm with [2], one of the most efficient multimedia retrieval algorithms. Here, the MinHash signature dimensionality is an important parameter which trades off between efficiency and accuracy. In this experiment, we adjust this parameter to ensure each retrieval algorithm achieve the best mAP. Then we compare their mAP, memory cost per video shot, and retrieval speed per query. The result is summarized in Table. 3. From the table, we could see that by incorporating temporal information into the inverted index, our retrieval algorithm achieves better accuracy and 9 times speedup but sacrifices memory efficiency. However, after combining scene descriptor with local feature, not only retrieval accuracy is improved from 0.8886 to 0.9401, but memory cost is reduced from 720 bytes to 480 bytes and retrieval speed is 17 times faster. This is because performing scene verification enables us to use MinHash signature of smaller dimensionality to represent each video.

6. Conclusion and future work

This paper has presented interest seam image, a novel approach to generating discriminant and efficient video synopsis for web-scale video content analysis applications, such as video recognition, video clustering, and video retrieval, etc.. A spatiotemporal energy map is defined to guide the extraction of prominent seams in the video, from which interest seam image is composited. Therefore, interest seam image preserves both spatially and temporally salient visual information in the videos. Its efficacy is demonstrated in a near duplicate web video retrieval task.

A novel video retrieval algorithm has been developed using interest seam image. It composes two novel components, i.e., an efficient inverted indexing scheme that takes advantages of the temporal context in the interest seam image, and a general post verification method, namely scene verification, which is manifested to be able to boost both retrieval accuracy and efficiency. Comparisons with state of the art video retrieval systems on a large scale web video database demonstrate that the proposed approaches simultaneously improves retrieval accuracy, retrieval speed, and memory efficiency.

Future works include further exploration of invariant visual representations for video which are robust to more types of video editings, and extensive tests in different types of video content analysis tasks and applications.

References

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *SIGGRAPH*, 2007.

[2] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *CIVR*, 2007.

[3] M. Douze, H. Jégou, H. Singh, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *ACM CIVR*, 2009.

[4] X. Zhang, Z. Li, L. Zhang, W. Ma and HY. Shum. Efficient Indexing for Large Scale Visual Search. In *ICCV*, 2009.

[5] J.Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[6] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[7] C. Kim and J.-N. Hwang. Object-based video abstraction for video surveillance systems. *IEEE TCSVT*, 2002.

[8] I. Laptev. On space-time interest points. *IJCV*, 2005.

[9] L. Liu, W. Lai, X.-S. Hua, and S.-Q. Yang. Video histogram: A novel video signature for efficient web video duplicate detection. In *MMM* , 2007.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. V. Gool. A comparison of affine region detectors. *IJCV*, 2005.

[12] C.-W. Nog, T.-C. Pong, and H.-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE TMM*, 2002.

[13] A. Oliva and A. B. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[15] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *SIGGRAPH*, 2008.

[16] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *ACM Multimedia*, 2009.

[17] H.-K. Tan, X. Wu, C.-W. Ngo, and W. Zhao. Accelerating near-duplicate video matching by combining visual similarity and alignment distortion. In *ACM Multimedia*, 2008.

[18] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *ACM Multimedia*, 2007.

[19] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1993.

[20] Y. Wexler and D. Simakov. Space-Time Scene Manifolds. *ICCV*, 2005.