Improving bag-of-visual-words model with spatial-temporal correlation for video retrieval.

WANG, L., SONG, D. and ELYAN, E.

2012

Improving Bag-of-visual-Words Model with Spatial-Temporal Correlation for Video Retrieval *

ABSTRACT

Most of the state-of-art approaches to Query-by-Example (QBE) video retrieval are based on the Bag-of-visual-Words (BovW) representation of visual content. It, however, ignores the spatial-temporal information, which is important for similarity measurement between videos. Direct incorporation of such information into the video data representation for a large scale data set is computationally expensive in terms of storage and similarity measurement. It is also static regardless of the change of discriminative power of visual words with respect to different queries. To tackle these limitations, in this paper, we propose to discover Spatial-Temporal Correlations (STC) imposed by the query example to improve the BovW model for video retrieval. The STC, in terms of spatial proximity and relative motion coherence between different visual words, is crucial to identify the discriminative power of the visual words. We develop a novel technique to emphasize the most discriminative visual words for similarity measurement, and incorporate this STC-based approach into the standard inverted index architecture. Our approach is evaluated on the TRECVID2002 and CC_WEB_VIDEO datasets for two typical QBE video retrieval tasks respectively. The experimental results demonstrate that it substantially improves the BovW model as well as a state of the art method that also utilizes spatial-temporal information for QBE video retrieval.

Categories and Subject Descriptors

H.3.4 [**Information Search and Retrieval**]: Query formulation,Retrieval models,Search process

General Terms

Theory

Keywords

Spatial-Temporal Correlation, discriminative visual word, Content based Video Retrieval, Query-by-Example, Bag-of-visual-Word

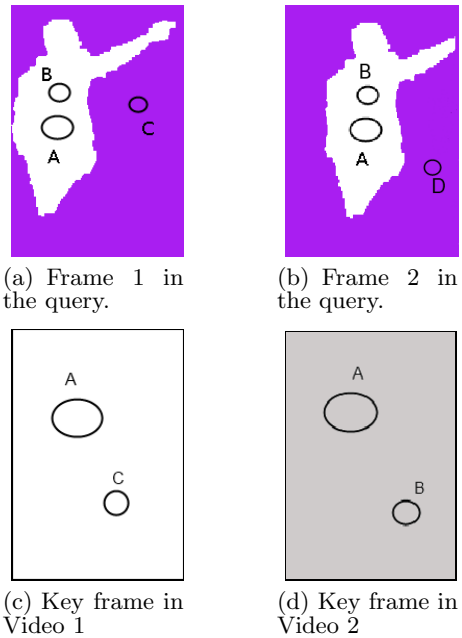
*

1. INTRODUCTION

The last decade has witnessed a rapid growth of video data over the Internet and online data repositories. There has been an urgent demand for more effective and efficient content-based video retrieval technology. In this paper, we focus on the Query-by-Example (QBE) video retrieval, which enables users to give a video example as the query to search against a data collection to find similar videos [26]. It is widely applied to many practical tasks such as near-duplicate video search, copyright infringement detection, instance search, etc.

Most state-of-art approaches to large scale content based visual retrieval are based on the Bag-of-visual-Words (BovW) model. As firstly introduced in the context of visual object search [23], high dimensional feature descriptors, such as SIFT [13] and SURF [1], are extracted to represent the stable and salient regions surrounding points-of-interest detected by local feature detectors. Popular detectors include Harris-Affine [14], Hessian-Affine [14], and Difference of Gaussian [13]. The regional descriptors generated from the collection or a training dataset are clustered and each cluster forms a visual word. Then given an image (or a keyframe in a video), the region descriptors in the image are quantized into discrete visual words. Specifically, the quantization function maps a region descriptor onto its closest cluster centroid. The region descriptor is then called an instance of the corresponding visual word (in this paper, we use the term "visual word" and "instance of visual word" interchangeably, for convenience, unless explicitly distinguished). As a result, an image can be represented as a bag of visual words. A pair of descriptors mapped onto an identical visual word are considered as a match between their visual contents. The similarity between two images can be measured based on the distance between their BovW representations.

In QBE video retrieval, a video is represented by a sequence of sampled key frames, each of which is represented by a bag (usually as a frequency histogram) of visual words. The similarity between two videos is obtained by aggregating the similarities of key frames across the two videos [4] . To alleviate the possible mismatches caused by unstable quantization, e.g. near identical regions being assigned to different visual words, soft-quantization of visual words [20, 8] has been proposed to map each descriptor onto multiple



(a) Frame 1 in the query.

(b) Frame 2 in the query.

(c) Key frame in Video 1

(d) Key frame in Video 2

Figure 1: Emphasizing the discriminative visual words to compensate the neglect of spatial-temporal internal structure. Query consists of two frames is represented by visual words A, B, C and D, and the video 2 is likely to be more relevant than 1)

neighboring visual words (in the descriptor feature space). Despite its simple structure, the BovW model has shown a promising performance in the fields such as object/event recognition [25] and image/video retrieval [21].

A major limitation of the BovW model, especially for QBE video retrieval, is that the spatial-temporal relation between visual words has been neglected, despite its obvious importance for similarity measurement of visual contents, mainly due to two reasons. First, the visual words within a frame are assumed independent of each other and the spatial relationship is discarded. Second, the temporal motions of visual words across the sequential frames are neglected.

Figure 1 shows an illustrative example. Here, we make a simplified assumption that the relevant videos should include the person appearing in the query. In Frame 1, the visual words A and B representing visual content of this person are messed up by another irrelevant visual word C. The BovW model, on one hand, discards the spatial relationship between A and B within the Frame 1, and on the other hand, neglects the motion of A and B from Frame1 to Frame 2. In the similarity measurement, the BovW model assumes that all visual words have an equal discriminative power. As a result, Video 1 and Video 2 are considered equally similar to the query. However, Video 2 is more likely to be relevant to the query, because the spatial relationship and motion coherence between A and B strongly implies that they belong to an identical object and should be more discriminative than C and D with respect to this query. Thus the spatial-temporal correlation of visual words A and B is a strong clue of relevance.

To overcome the limitations, various approaches (described in more detail in Section 2) have been proposed to incorporate the spatial-temporal constraints associated with visual words. Some recent image/video retrieval methods add position, scale, main orientation and motion primitives of each visual word directly into the BovW representation, and then, for example, enhance the similarity measurement with Weak/Tight Geometric Constraint (WGC/TGC) verification of matched visual words between two images/videos [7, 32]. Nonetheless, the injected information results in a significant increase of computation cost in the similarity match. Another direction is to expand the vocabulary with spatially correlated visual word combinations [30, 29, 33] to form "visual phrases". However, it tremendously increases the storage cost of visual content representation, and does not take into account the temporal (motion) constraint.

To tackle the aforementioned limitations, instead of expanding or adding extra spatial-temporal information directly into the BovW representation, we propose to identify and emphasize the most discriminative visual words (for example, A and B in the Figure 1) through exploiting the spatial-temporal correlation in an integrated manner, among different visual words in the query. The emphasizing of discriminative visual words would revise the query representation and exclude the irrelevant information, which compensates the neglect of spatial-temporal information. Furthermore, the spatial-temporal information discovery from query example does not result in extra storage cost for data representation nor increased complexity in similarity measurement.

We propose the characterize the discriminative visual words based on spatial proximity and temporal motion coherence. In the consecutive frames, an inherent object often has an explicit spatial structure and intensive spatial relationship. The motion of object layout across neighboring frames in a video often has a characteristic of coherence. This spatial-temporal relationship can be utilized to approximate the possibility that the visual words are associated with an identical object, and such visual words usually have more discriminative power. In this paper, we base our proposed method on two assumptions on the discriminative visual words: i) they co-occur closely in a frame; ii) they move coherently across sequential frames. The spatial proximity and motion coherence is termed as spatial temporal correlation (STC).

In this paper, we present a novel technology to model the STC imposed by a query. Specifically, we propose to model the pair-wise STC by a Gaussian distribution based on the assumption that the layout of objects follows the law of Mixture Gaussian distribution [2, 12]. A similarity measure based on STC is developed, which emphasizes the discriminative visual words with respect to the query. Essentially, this leads to reformulation of the key frames of the query video, effectively involving the discriminative visual words that may or may not originally appear in the key frame and excluding the noisy ones. The discriminative power of a visual word are determined by both the STC matrix and their frequencies. Furthermore, it is important to note that the retrieval technology can be easily incorporated into standard inverted indexing architecture to achieve high computational efficiency.

The rest of the paper is organized as follows: Section 2 reviews the related work on incorporating spatial-temporal information in content based image/video retrieval; Section 3 describes our STC-based video retrieval technology in detail; Section 4 presents the experimental settings and results for an extensive evaluation on two typical QBE video retrieval tasks; Section 5 concludes the paper and points out future research directions.

2. RELATED WORK

There have been some recent approaches that utilize the spatial or temporal constraint associated with the visual words to improve the BovW model. For example, spatial information is introduced into image/video retrieval [18, 19, 23] for a post-retrieval re-ranking, which matches visual words through verification of their neighboring visual words between two images. Sivic et al. [23] employed weak temporal constraint to remove the visual words which can not be tracked in 3 consecutive frames of video. Chum et al. [5] proposed a generative model to expand the query based on the strong spatial constraints discovered from the first round retrieval results and the query for image retrieval.

Utilizing additional spatial-temporal information incorporated in the video representation, the idea of the WGC/TGC [7, 32] is that the scales and the main direction variations of the correctly matched visual words should be consistent. Spatial Pyramid Matching [11] is proposed to approximate the global geometric correspondence. It partitions the image into hierarchical sub-images to model the spatial relation of visual words and improve the similarity measurement.

Machine learning based technologies have been proposed to join related visual words into visual phrases according to the spatial constraint. Sivic et al. [22] applied the latent semantic indexing based on co-occurring visual words. Zhang et al. [29] developed a supervised learning technique to combine co-occurring pairs of visual words to representative visual phrases for known categories of images. Zhang et al. [30] proposed an approach to quantize an image into bins and encode the spatial information with the co-occurred visual words as geometry-preserving visual phrases to achieve higher discriminative power. However, these approaches largely increase the dimensionality of visual content representation. Temporal information is commonly used in the recent video analysis technology based on the BovW model. For example, Wang et al.[25] proposed to incorporate a number of motion primitives of each visual word into the BovW representation of videos. Nevertheless, the above approaches usually directly include the spatial-temporal information into visual content representation, and the storage and computational cost is often high. This is one of the major obstacles for developing large scale content-based visual information retrieval technology. Furthermore, the spatial-temporal information incorporated in video representation is also not straightforward to be deployed with inverted index architecture.

There have been approaches to deriving the discriminative features according to spatial-temporal constraints at the pixel level for image/video retrieval. For example, similar to spatial points-of-interest detection, various spatial-temporal local points-of-interests detectors are proposed, such as Spatial-Temporal Interest Points [10, 2] and Volumetric features [9]. The spatial-temporal local detectors are designed to detect the salient "spatial-temporal corners" or "sub-volume" of a video at pixel level, which has succeeded in action or event recognition. Nonetheless, at the pixels level it would not be able to capture the spatial-temporal constraints from a "semantic" perspective. Furthermore, there are attempts to enhance the BovW by utilizing higher level spatial-temporal constraints between multiple visual words. Object-of-Interest extraction according to spatial-temporally distribution of the visual words are proposed [12], and it improves the visual object search. However, the Object-of-Interest based representation would lose much information associated with the background of videos. Overall, these are interesting directions for further exploration but out of scope of this paper, where we focus on the visual words level in line with the state of the art in QBE video retrieval.

3. VIDEO RETRIEVAL BASED ON STC

3.1 Video Representation based on the BovW Model

In this paper, each video in a collection is represented as a set of key frames $v_d = \{\boldsymbol{f}_d\}$, where each key frame $\boldsymbol{f}_d$ is represented by a K dimensional vector (K denotes the size of vocabulary) of visual words occurring in the frame: $\boldsymbol{f} = \{w_i\}$ where the weight w_i of the i^{th} visual words in $\boldsymbol{f}$ is its Term Frequency (TF). Similarly, a given query example, which is also a video, is represented as $v_q = \{\boldsymbol{f}_l\}$ where f_l is a frame. For efficiency, a bunch of key frames $\{\boldsymbol{f}_q\} \subset v_q$ is sampled for the video similarity measurement. Note, however, that we use all frames $\{\boldsymbol{f}_l\}$ in the query for spatial-temporal correlation detection and measurement as shown in the next subsections.

The BovW model usually involves a very large vocabulary, and the representation vector $\boldsymbol{f}$ is sparse. Therefore, the inverted index architecture can be applied. For each visual word, a table is built to list all the frames where it appears. The key frame similarity $sim(\boldsymbol{f}_d, \boldsymbol{f}_q)$ is measured by the cosine function, which can be approximated by:

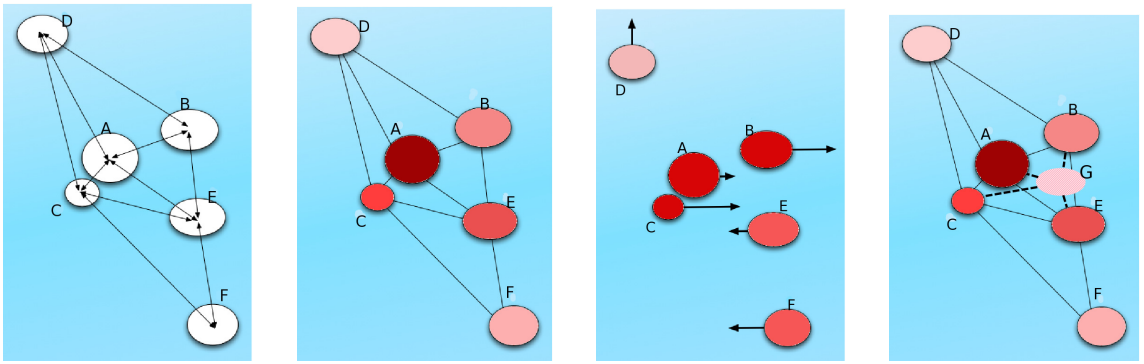
$$sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \approx \frac{\sum_{i=1}^{K} score(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{1}$$

where scoring function $score(w_i)$ is defined as:

$$score(w_i) = \boldsymbol{f}_q(w_i) * \boldsymbol{f}_d(w_i) \tag{2}$$

where $l(\boldsymbol{f})$ is the L^2-Norm of vector $\boldsymbol{f}$, and $score(w_i)$ is the scoring function of each matched visual word across $\boldsymbol{f}_d$ and $\boldsymbol{f}_q$, given by the multiplication of the corresponding TFs. The scores are accumulated to compute a similarity score between two key frames. In this paper, it is assumed that the video data are well segmented shots or the videos are short and consist of few shots. We then adopt the shot similarity measurement method proposed by Peng et al [17] where the highest similarity score among all possible pairs of key frames compared is used to measure the similarity between two video shots:

$$sim_{v_d, v_q} = \max_{\boldsymbol{f}_d \in v_d, \boldsymbol{f}_q \in v_q} sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \tag{3}$$



(a) Visual words in the key frame of query video

(b) Spatial correlation based emphasizing

(c) Motion coherence based emphasizing

(d) Emphasizing the discriminative visual word not appearing in the key frame

Figure 2: An example of visual words emphasizing approach. In (b), (c) and (d), the color intensity indicates the importance.

3.2 Characterizing Discriminative Visual Words

We aim to improve the BovW model by emphasizing the visual words satisfying the spatial proximity and temporal motion coherence constraints. Effectively, these visual words tend to be associated with an identical object. Traditionally, this association was addressed by image segmentation, which is computationally expensive [6]. Reliable segmentation and acceptable segmentation quality are also open questions. In this paper, we approximate the discriminative power of visual words according to the two assumptions proposed in Section 1. We develop a novel characterizing scheme by considering its spatially and temporally correlated visual words.

Regarding the spatial correlation, it can be measured by the proximity between visual words, e.g. the inverse of Euclidean Distance [28, 12]. An example is illustrated in Figure 2(a). Visual word A is located in a close proximity to B, C and E in the key frame of the query, and visual word A is assigned with a higher discriminative power as shown in Figure 2(b).

In addition, in the temporally consecutive frames, the reoccurring visual words are tracked according to the L^2 Norm. Each tracked visual word moves to new position in the next frame, and associated with a motion vector (Figure 2(c)). We propose to measure the relative motion with respect to two visual words to indicate how the two visual words move coherently. In Figure 2(c), visual words A, B, and C move coherently to right, and they are assigned with a higher discriminative power.

Eventually, if a visual word had a strong STC with other words appearing in the same frame of query, the visual word is assumed to have a higher discriminative power.

Based on the above principles, our scoring function based on STC for a visual word in a key frame of the query video is then formulated as follows:

$$score'(w_i) = \sum_{j=1}^{K} \boldsymbol{f}_q(w_i) * st(i,j) * \boldsymbol{f}_d(w_j) \tag{4}$$

where $st(i,j)$ measures the STC between the i^{th} and j^{th} visual words. The formulation of STC measurement function will be shown in the next subsection.

3.3 The STC Measurement Function

As discussed in previous section, the STC measurement is formulated based on the principles shown in Figure 2. The STC between a pair of visual words should be inversely proportional to the physical space distance and relative motion associated.

As has been shown in [12, 3], the visual objects' spatial layout can be modeled by a Gaussian Mixture Model. The spatial correlation between pair-wise visual words in a frame of the query video is then formulated as follows:

$$s_{i,j} = \begin{cases} \sum_{a=1}^{\boldsymbol{f}_q(w_i)} \sum_{b=1}^{\boldsymbol{f}_q(w_j)} e^{-\kappa d_{a,b}^2} & \text{for} \quad \boldsymbol{f}_q(w_i) \& \boldsymbol{f}_q(w_j) \neq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

where $d_{a,b} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$ denotes the Euclidean distance between the a^{th} instance of the i^{th} visual word and the b^{th} instance of the j^{th} visual word in the frame, as shown in Figure 3. It should be noted that in a frame represented by a very large vocabulary, there usually exists only one instance for each visual word, and the spatial correlation $s_{a,b}$ between the two instances directly represents the spatial correlation between the two visual words. The parameter κ analogs the inverse of variance in the Gaussian Distribution, which controls the decreasing speed of the spatial correlation. Generally, the larger is the parameter κ, the more visual words are considered as spatial correlated.

Similar to the spatial correlation measurement, the modeling of temporal motion coherence is also based on the Gaussian

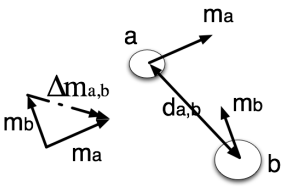


Figure 3: Relative Motion between a pair of instances of visual word

distribution. The measurement function is formulated by:

$$t_{i,j} = \begin{cases} \sum_{a=1}^{\boldsymbol{f}_q(w_i)} \sum_{b=1}^{\boldsymbol{f}_q(w_j)} e^{-\gamma \|\Delta \boldsymbol{m}_{a,b}\|^2} & \text{for} \quad \boldsymbol{f}_q(w_i) \& \boldsymbol{f}_q(w_j) \neq 0 \\ 0 & otherwise \end{cases} \tag{6}$$

where $\Delta \boldsymbol{m}_{a,b}$ denotes the relative motion between instance a of the i^{th} visual word and instance b of the j^{th} visual word in the frame. The relative motion vector $\Delta \boldsymbol{m}_{a,b}$ is calculated by:

$$\Delta \boldsymbol{m}_{a,b} = \boldsymbol{m}_a - \boldsymbol{m}_b \tag{7}$$

where $\boldsymbol{m}$ denotes the motion vector of the tracked instance of the visual word in consecutive frames, as shown in Figure 3. Similarly, the parameter γ in Equation 6 controls the decreasing of temporal correlation along with the expansion of the relative motion.

Equations 6 and 5 are combined to model the STC for a pair of visual words. To simplify the computation, an addition is utilized to fuse the two correlations. If any visual word in the pair is not tracked in next frame, the temporal correlation is set to zero. The STC within a frame and the entire query video are calculated as:

$$\begin{aligned} st^l_{i,j} &= (s^l_{i,j} + t^l_{i,j}) \\ ST_{i,j} &= \sum_{l=1}^{L} st^l_{i,j} \end{aligned} \tag{8}$$

where $st^l_{i,j}$ represents the STC between the i^{th} and j^{th} visual words in the l^{th} frame of query video v_q, and L denotes the total number of frames in v_q. $ST_{i,j}$ is the STC between the i^{th} and j^{th} visual words measured based on the entire query. In this way, the STC of singular and rarely appearing visual words are relatively reduced.

In summary, an STC matrix ST is generated for the query, and each entry of ST is a pair-wise correlation computed by Equation 8. How the STC incorporated to improve the BovW based similarity measurement will be described in the next subsection.

3.4 Key Frame Reformulation and Similarity Measurement

By incorporating STC-based scoring function (Equation 4), the key frame similarity measure in Equation 2 becomes:

$$sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \approx \frac{\sum_{i=1}^{K} score'(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{9}$$

However, direct computation of Equation 9 is difficult to implement for the inverted index architecture due to the STC computation. The original scoring function of Equation 2 corresponds to the inner product between the query and data representation vectors, which can be easily applied to inverted index system. To facilitate a similar computation, the numerator of Equantion 9 is rewritten as follows:

$$\begin{aligned} \sum_{i=1}^{K} score'(w_i) &= \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{f}_d(w_j) * \boldsymbol{f}_q(w_i) * st(i,j) \\ &= \sum_{j=1}^{K} \boldsymbol{f}_d(w_j) * \sum_{i=1}^{K} st(i,j) * \boldsymbol{f}_q(w_i) \\ &= \sum_{j=1}^{K} \boldsymbol{\alpha}(w_j) * \boldsymbol{f}_d(w_j) \end{aligned} \tag{10}$$

The weighting vector $\boldsymbol{\alpha} = ST \times \boldsymbol{f}_q$ and $\alpha \in \mathbb{R}^K$, where $\alpha(w_i)$ denotes the discriminative power (also called emphasizing weights) of the i^{th} visual word, and can also be directly used as the emphasizing weights. ST is the $K \times K$ STC matrix computed by Equation 8.

As shown in Equation 10, the STC is injected into the scoring function for emphasizing the discriminative visual words in the keyframes, which is equivalent to key frame reformulation. It must be noted that the STC is measured using all frames in the query (Equation 8), and as a result, some visual words may be brought by the STC into the reformulated key frame, even though their original term frequency on the key frame is zero. For example, as shown in Figure 2(d), the visual word G is added with corresponding weights, because of its strong STC with A, B, C and E in the whole query. In this way, the STC based approach would, to some extent, compensate the information loss caused by the key frame sampling.

Equation 11 has been utilized to further quantize the emphasizing weights of the discriminative visual words in the key frame. This is to avoid the risk of assigning an extreme weight to some individual visual word:

$$\boldsymbol{\alpha}'(w_i) = \begin{cases} 2 & \text{for} \quad \boldsymbol{\alpha}(w_i) > \sigma \\ 1 & \text{for} \quad \sigma/2 < \boldsymbol{\alpha}(w_i) < \sigma \\ 0 & \text{for} \quad else \end{cases} \tag{11}$$

As shown in Figure 4, the choice of σ determines the number of discriminative visual words. Figure 4 represents some examples of the $\alpha'(w_i)$ computed for various queries. Given the same detector is used, the discriminative power distributions for different queries are not very different. Thus a static threshold σ can be empirically selected for all queries (σ=1 in our experiment). The effect of quantization scheme will be discussed in Section 4.

Based on Equations 9 and 10, the frame level similarity measurement function becomes:

$$sim'(f_d, f_q) \approx \frac{\sum_{i=1}^{K} \boldsymbol{\alpha}'(w_i) * \boldsymbol{f}_d(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{12}$$

It is important to note that the key frame reformulation will not largely increase the number of non-zero elements in the key frame representation. It would not only involve the discriminative visual words, but also exclude the noises. Furthermore, it avoids the extra computational and memory

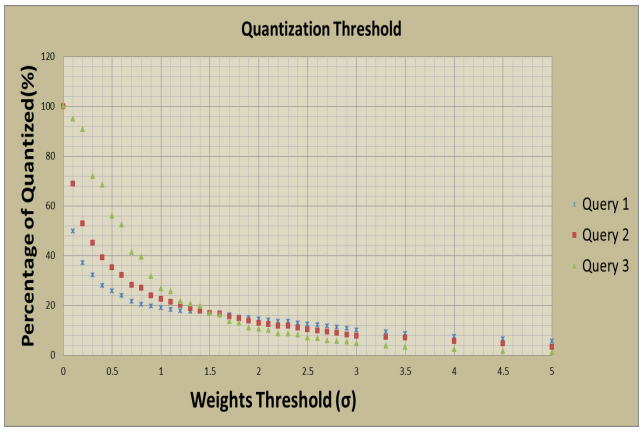


Figure 4: The quantization threshold σ decides the number of emphasized visual words.

costs of the direct inclusion of the spatial-temporal information into video representation and indexing.

Having said that, in this paper, we are more interested in how the STC-based approach can improve the retrieval effectiveness. In the next section, we present an extensive empirical evaluation.

4. EXPERIMENTS

The goal of these experiments is to evaluate the effectiveness of the STC-based approach to improve the BovW model. In general, the video relevance can be defined on two levels: the visual level and concept level. Accordingly two Query-by-Example video retrieval tasks are used: (1) QBE near-duplicate video search task for searching visually similar videos; (2) general topics QBE video retrieval task, for retrieving "conceptually" similar videos.

In our experiments, salient points are detected by the Hessian detector, which works well to overcome the occlusion and cluttering [15]. The salient regions are described by SIFT feature. Hierarchical K-means [16] is used for visual vocabulary construction.

Mean Average Precision (MAP) is used as main performance indicator, and we also show the Precision-Recall curves of different models.

The classical BovW model and a state of the art BovW enhancement approach based on Tight Geometric Constraint (TGC) [32] are used as the baselines. The TGC method is implemented with a publicly available toolkit SOTU [31]. The performances of four variations of the similarity measurement approaches proposed in this paper are reported, namely the STC-based retrieval models with and without weights quantization (Equation 11), denoted by **st-BovW** and **raw-st** respectively; and retrieval model based on spatial (temporal) correlation only, denoted by **s-BovW** (**t-BovW**).

4.1 Experimental Set Up

Two datasets commonly used datasets are selected for the experiments:

(1) `CC_Web_Video` Near-duplicate video search is performed on data collection CC_Web_Video [27]. Most videos in this data-collection are short videos which are mostly 3-5 minutes long and not longer than 10 minutes. They are presented on the website of Youtube, Yahoo and Google Video. From original data collection, we randomly selected 4590 videos to form the experimental data collection, and totally 336K key frames are extracted to represent the videos in the data collection. The videos in the ground truth are labeled by "*Exact duplicate*", "*Similar*", "*Major Changed*", "*Long version*" and "*Not Relevant*". The evaluation is performed for 24 topics, using 69 queries respectively selected from "*Exact*", "*Similar*" and "*Major Changed*" videos. 10 key frames are sampled from each query for retrieval. The average number of the relevant videos is 84.7 per topic.

(2) `TRECVID2002` The general topic QBE video retrieval task is performed on the dataset TRECVID2002 [24]. This dataset contains various video sources: old film, news, documentary and advertisement. The retrieval task consists of 7 topics and each topic has at least 3 different queries. The topics involve searching for specific person, object, action, scene or instances of a category of person. The data collection is composed of 40 randomly selected videos, which are segmented to approximately 3000 shots based on the shot boundary data provided by TRECVID2002. Each shot is normally no longer than 1 minute and associated with a single scene or single semantic concept. Totally, 30K keyframes are extracted. On average, each topic is associated with 23 relevant shots.

4.2 Experiment 1: QBE Near-Duplicate Video Search

As shown by the Precision-Recall curves in Figure 5, the st-BovW outperforms the classical BovW and TGC methods. The raw-st performs similarly to the s-BovW, t-BovW and st-BovW.

Figure 6(a), 6(b) and 6(c) present the performances of different approaches using "Exact", "Similar" and "Major Changed" queries respectively. The "Exact" queries have the highest quality, and the quality of "major changed" is the lowest. The influence of query quality on the performance of QBE video retrieval is obvious. The performance of good queries ("Exact" and "Similar") is a lot better than the other type of queries.

Furthermore, as shown in Figure 6(a) and Figure 6(b), the classical BovW performs adequately well on the "Exact" and "Similar" queries. However, for "Major Changed" queries, which are of low quality, as shown by the Precision-Recall curves in Figure 6(c), both the st-BovW and TGC approaches outperform the classical BovW model, and the st-BovW outperform TGC.

The performances in term of MAP are presented in Table 1, where E, S, M denotes "Exact", "Similar" and "Major Changed" queries, respectively. On average, the st-BovW outperforms the classical BovW by 4%, which is statistically significant (p-value=0.046). The st-BovW performs much better than TGC for the low quality queries and slightly better for higher quality queries. The st-BovW outperforms TGC by about 10% and BovW by about 20% on the "Major

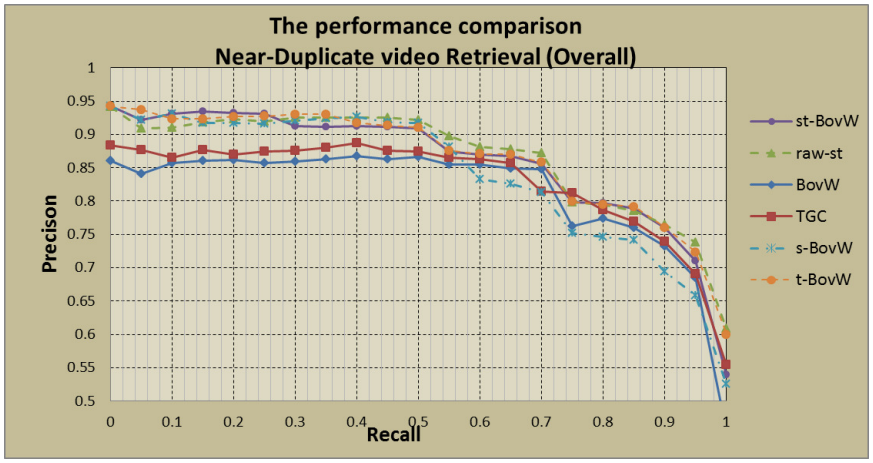
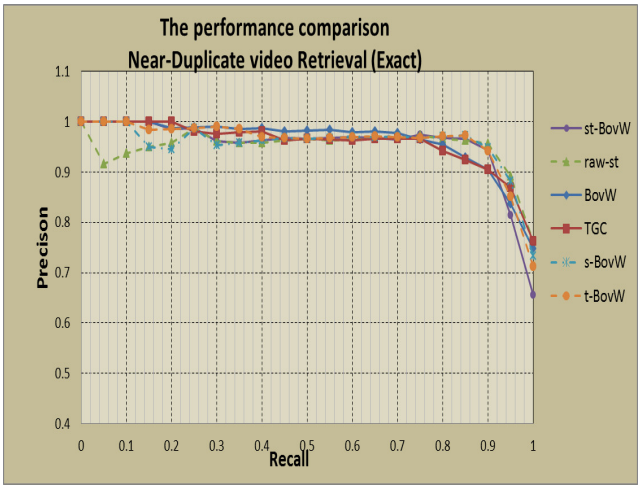
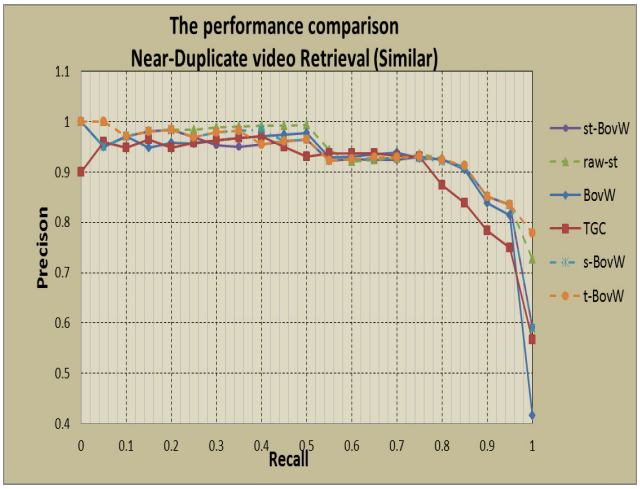


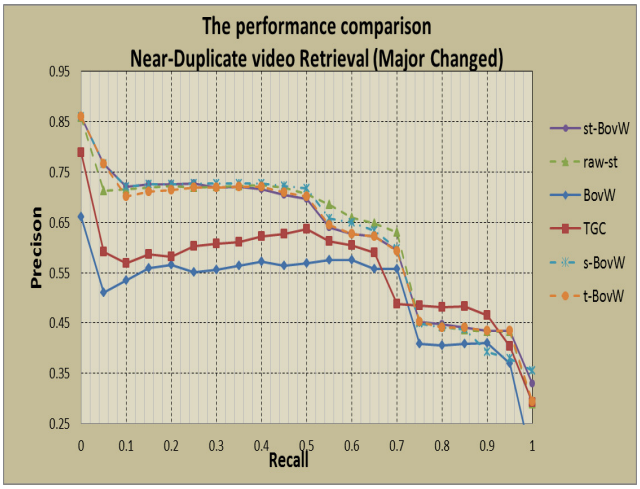
Figure 5: Precision-Recall curve for overall queries



(a)



(b)



(c)

Figure 6: Query by video examples labelled by (a) "Exact", (b)"Similar", (c)"Major Changed"

Table 1: Mean Average Precision Comparison

MAP	E	S	M	Overall
BovW	0.939	0.919	0.511	0.8170
TGC	0.940	0.913	0.553	0.8268(+0.010)
st-BovW	**0.944**	0.927	0.619	**0.8503(+0.033)**
raw-st	0.929	**0.930**	**0.625**	0.8477(+0.030)
s-BovW	0.938	0.925	0.619	0.8360(+0.021)
t-BovW	0.941	0.926	0.615	0.8486(+0.032)

Figure 8: Typical frames of relevant videos of topic "Women in long dresses"

Changed" queries.

Noted that the raw-st performs nearly the same as the st-BovW, s-BovW and t-BovW in terms of Precision-Recall curves, and in term of MAP on average. However, as shown in Figure 5 and Figure 6(a), the precision of top ranked results retrieved by raw-st is not as good as other approaches. This may be owing to the extreme weights assigned to some visual words which may be noise in some cases. The weights quantization scheme effectively reduces the risk of extreme weights for the discriminative visual words generated by the STC-based key frame reformulation and make the st-BovW perform more stably.

In summary, the evaluation on the near duplicate video search task shows that both TGC and st-BovW effectively improve the performance of the BovW model. st-BovW largely outperforms TGC on lower quality queries, which is more challenging for classical BovW model, and it has also performed comparably to TGC on high quality queries.

4.3 Results: General Topics QBE Video Retrieval

The QBE video retrieval for general topics is always a challenging task. The topics cover the various user intentions, who may search for specific object/scene or a category of shot. The desired object may appear visually different in the relevant videos. For example, as shown in Figure 8, one of the topics in this experiment is searching for videos that contain women wearing long dresses. In the relevant videos, the persons may appear differently, and only a small amount of visual content maintained by the relevant videos are visually matching the query. In other words, there is semantic gap between the visual similarity and topic relevance.

In this experiment, we use a vocabulary of a small size 5000. The goal is to evaluate if the proposed STC-based discriminative visual words emphasizing technique can effectively compensate the neglect of spatial-temporal structure in the BovW model.

Table 2: Average Precision Comparison

AP	BovW	TGC	st-BovW	raw-st
Eddie Rickbacker	0.1185	0.1119	0.1372	**0.1510**
Musician playing music instruments	0.1257	0.1083	**0.1602**	0.1490
Women in long dresses	0.0816	0.0851	0.0960	**0.1009**
Overhead Views of Cities	0.1204	0.1096	**0.1302**	0.1162
Oil fields, rig and Oil Equipment	**0.1586**	0.1500	0.1411	0.1414
Map of the Continental	0.1508	**0.1638**	0.1574	0.1243
Live beef or Dairy Cattle	0.099	0.0845	**0.1156**	0.1132
Overall	0.1220	0.1162	**0.1340**	0.1280

Table 3: Mean Average Precision Comparison between STC based Approaches

Approaches	s-BovW	t-BovW	st-BovW	raw-st
MAP	0.1258	0.1140	**0.1340**	0.1280
Average Precision on 10% Recall	0.528	0.458	**0.610**	0.466

Figure 7 demonstrates that st-BovW outperforms the classical BovW model and TGC model. Table 2 presents the Average Precisions of different approaches on individual topics. It is shown that st-BovW outperforms the classical BovW model on 6 out of 7 topics. The average improvement is 10% and it is statistically significant (p-value=0.044). The experimental results prove that the discriminative visual words emphasizing technique based on STC also improves the performance of BovW model using relatively small vocabulary.

It is also shown in the Table 2, that st-BovW outperforms TGC on 5 topics, except for two topics: "Oil fields, rig and Oil Equipment" and "Map of the Continental". On the other hand, TGC fails in improving the classical BovW model in other 4 topics.

As can be seen in Table 3, the spatial correlation seems to become more crucial than temporal correlation. The evidence is that s-BovW outperforms t-BovW on average. In the query video for a general topic, the visual contents representing users' desire are often messed up with large amount of noise. Thus the emphasizing technique based on the spatial proximity assumption is more effective than the motion coherence to reduce the noise.

The weights quantization scheme has improved the performance of STC-based approaches as shown in Table 2. ST-BovW performs more stably than raw-st, and raw-st performs lower than the BovW baseline on 3 topics although achieving highest performance on other two topics. Another

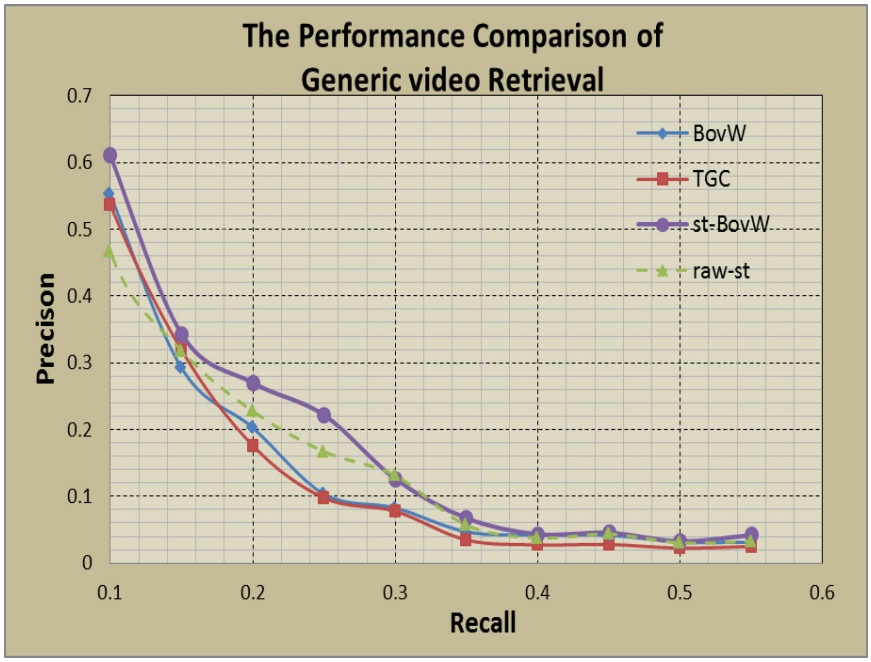


Figure 7: Precision-Recall curve for queries with different key-frames

evidence is shown in Table 3, and the weights quantization scheme has promoted the precision on the top ranked retrieved results.

In summary, the evaluation demonstrates that STC based technique can effectively improve the performance of classical BovW, and st-BovW performs more stably than TGC-based approach on the general topic QBE video retrieval.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel approach based on the spatial temporal correlation (STC) for QBE video retrieval. The STC is discovered from the query and it is used to address the discriminative visual words in the BovW similarity measurement. Furthermore, we propose to emphasize the discriminative visual words and reformulate the key frames in the query. The similarity measurement function can be easily implemented based on the quantized emphasizing weights.

A series of experimental results on the near-duplicate web video search and general topic video retrieval tasks show that the STC-based approach substantially improves the classical BovW model without increasing storage cost for video representation. The STC-based approach has also outperformed the state of the art TGC-based approach on some challenging tasks. The results have verified our hypothesis that the discriminative visual words can be characterized and emphasized according to STC to effectively compensates the neglect of spatial-temporal information and the information loss during the key-frame sampling in the classical BovW model.

A possible future research direction is to investigate in developing new data representation model based on the discriminative visual words according to the spatial-temporal correlation discovered from the whole video collection.

6. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[2] L. Cao, Y. Tian, Z. Liu, B. Yao, Z. Zhang, and T. S. Huang. Action detection using multiple spatial-temporal interest point features. In *ICME*, pages 340–345, 2010.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8):1026 – 1038, aug 2002.

[4] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 549–556, New York, NY, USA, 2007. ACM.

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[7] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *Int. J. Comput. Vision*, 87:316–336, May 2010.

[8] Y.-G. Jiang and C.-W. Ngo. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Comput. Vis. Image Underst.*, 113:405–414, March 2009.

[9] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 166–173, Washington, DC, USA, 2005. IEEE Computer Society.

[10] I. Laptev and T. Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[12] D. Liu and T. Chen. Video retrieval based on object discovery. *Comput. Vis. Image Underst.*, 113:397–404, March 2009.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[14] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, Oct. 2004.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, Nov. 2005.

[16] D. NistŐr and H. StewŐnius. Scalable recognition with a vocabulary tree. In *IN CVPR*, pages 2161–2168, 2006.

[17] Y. Peng and C.-W. Ngo. Emd-based video clip retrieval by many-to-many matching. In *CIVR*, pages 71–81, 2005.

[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR 2007. IEEE Conference on*, june 2007.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Imprfoving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[21] H. Ren, S. Lin, D. Zhang, S. Tang, and K. Gao. Visual words based spatiotemporal sequence matching in video copy detection. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, pages 1382–1385, Piscataway, NJ, USA, 2009. IEEE Press.

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 370 – 377 Vol. 1, oct. 2005.

[23] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006.

[24] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[25] F. Wang, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and visual relatedness. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 239–248, New York, NY, USA, 2008. ACM.

[26] L. Weng, Z. Li, R. Cai, Y. Zhang, Y. Zhou, L. T. Yang, and L. Zhang. Query by document via a decomposition-based two-level retrieval approach. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 505–514, New York, NY, USA, 2011. ACM.

[27] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan. Real-time near-duplicate elimination for web video search with content and context. *Trans. Multi.*, 11:196–207, February 2009.

[28] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, page 13, june 2006.

[29] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 75–84, New York, NY, USA, 2009. ACM.

[30] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

[31] W.-L. Zhao. http://www.cs.cityu.edu.hk/ wzhao2/ sotu.htm.

[32] W.-L. Zhao, X. Wu, and C.-W. Ngo. On the Annotation of Web Videos by Efficient Near-Duplicate Search. *IEEE Transactions on Multimedia*, 12(5):448–461, Aug. 2010.

[33] Q.-F. Zheng and W. Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:7:1–7:19, October 2008.