12-2009

Coherent phrase model for efficient image near-duplicate retrieval

Yiqun HU

Xiangang CHENG

Liang-Tien CHIA

Xing XIE

Deepu RAJAN

See next page for additional authors

Coherent Phrase Model for Efficient Image Near-Duplicate Retrieval

Article *in* IEEE Transactions on Multimedia · December 2009

DOI: 10.1109/TMM.2009.2032676 · Source: DBLP

6 authors, including:

Yiqun Hu

42 PUBLICATIONS **1,386** CITATIONS

SEE PROFILE

Liang-Tien Chia

Nanyang Technological University

157 PUBLICATIONS **3,520** CITATIONS

SEE PROFILE

Ah-Hwee Tan

Nanyang Technological University

234 PUBLICATIONS **4,059** CITATIONS

SEE PROFILE

Coherent Phrase Model for Efficient Image Near-Duplicate Retrieval

Yiqun Hu, Xiangang Cheng, Liang-Tien Chia, *Member, IEEE*, Xing Xie, *Senior Member, IEEE*, Deepu Rajan, *Member, IEEE*, and Ah-Hwee Tan, *Senior Member, IEEE*

***Abstract*—This paper presents an efficient and effective solution for retrieving image near-duplicate (IND) from image database. We introduce the coherent phrase model which incorporates the coherency of local regions to reduce the quantization error of the bag-of-words (BoW) model. In this model, local regions are characterized by visual phrase of multiple descriptors instead of visual word of single descriptor. We propose two types of visual phrase to encode the coherency in feature and spatial domain, respectively. The proposed model reduces the number of false matches by using this coherency and generates sparse representations of images. Compared to other method, the local coherencies among multiple descriptors of every region improve the performance and preserve the efficiency for IND retrieval. The proposed method is evaluated on several benchmark datasets for IND retrieval. Compared to the state-of-the-art methods, our proposed model has been shown to significantly improve the accuracy of IND retrieval while maintaining the efficiency of the standard bag-of-words model. The proposed method can be integrated with other extensions of BoW.**

***Index Terms*—Bag-of-word (BoW), image near-duplicate (IND), quantization, retrieval, TRECVID.**

I. Introduction

RETRIEVAL and detection of image near-duplicate (IND) [1], [2] are very useful for the filtering, retrieval, and management of multimedia contents. For example, the INDs can correlate the videos that depict the same news event from different broadcast sources and provide similarity clues for recognizing visual events and searching news video clips [3]. Detecting IND(s) over the Internet can discover unauthorized use of private images for the application of copyright infringement detection [3], [4]. Personal photo album can be automatically organized by grouping/removing IND(s), which might be of different names. Detection and retrieval of IND can also facilitate traditional text-based web search. If two web pages contain any IND(s), the relevance rate between these two web pages should be increased.

Manuscript received September 02, 2008; revised May 18, 2009. First published September 22, 2009; current version published November 18, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale.

Y. Hu, X. Cheng, L.-T. Chia, D. Rajan, and A.-H. Tan are with the School of Computer Engineering, Nanyang Technological University, 639798 Singapore (e-mail: yqhu@ntu.edu.sg; xiangang@pmail.ntu.edu.sg; asltchia@ntu.edu.sg; asdrajan@ntu.edu.sg; asahtan@ntu.edu.sg).

X. Xie is with Microsoft Research Asia, Beijing 100080, China (e-mail: xingx@microsoft.com).

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

Digital Object Identifier 10.1109/TMM.2009.2032676

Fig. 1. Four IND pairs in the dataset provided by [1] which is successfully detected by the proposed method (shown in the second row) but failed by [2] (shown in the third row).

According to [1], IND is referred to as multiple images that are close to the exact duplicate of one image, but different in the composition of the scene, camera setting, photometric, and digitization changes. Specifically, the scale, viewpoint, and illumination of the same scene and object(s) captured in the IND(s) can be changed by different camera settings and rendering conditions. The composition of multiple objects can be different in the IND(s) due to some editing operations. Fig. 1 shows four examples of INDs with various differences: There are scale changes in the two images of the first two IND pairs as shown in the first two columns. The two images of the second and fourth IND pairs shown in the corresponding columns are different due to viewpoint change. The lighting conditions of the two images are different for the first three IND pairs. Due to editing operation, the compositions of two persons are different in the two images of the third IND pair.

Retrieval and detection are two different but related tasks about IND. IND retrieval aims to find all images that are duplicate or near duplicate to a query. The objective of IND detection is to find all duplicate pairs from an image collection. IND detection is based on IND retrieval but is more challenging than IND retrieval because of two reasons: 1) there are n^2 image pairs needed to be checked in a dataset of n images; 2) besides calculating similarity, the detection algorithm needs to decide whether to accept as IND or reject. There are two issues related to IND detection and retrieval: 1) the large variances within INDs make this problem challenging; 2) by formulating detection as a retrieval problem, the number of possible IND pair increases quadratically with the size of the database. It leads to the deterioration in the efficiency of retrieval. So the solution should be both *robust* and *efficient* in comparing a pair of images to determine whether or not they are IND.

In this paper, we propose an efficient and effective coherent phrase model (CPM). This model uses **visual phrase** of multiple descriptors instead of **visual word** of single descriptor to characterize every local region to enforce local coherency. Specifically, we propose two types of phrase.

- Feature coherent phrase (FCP): every local region is characterized by multiple descriptors of different types, e.g., SIFT [5] SPIN [6]. The match of two local regions requires the coherency across different types of feature.
- Spatial coherent phrase (SCP): multiple descriptors of a single type of feature are generated from the spatial neighborhoods of different scales around every local region. The match of two regions requires the coherency across different spatial neighborhoods.

The proposed coherent phrase model has several advantages. First, it improves the *effectiveness* of bag-of-words (BoW) model by introducing visual phrase, which enforces the coherency across multiple descriptors to remove false matches. The quantization error introduced by visual codebook generation [7], [8] can be reduced by such local coherency information. Second, it provides an *efficient* framework to explore appearance/spatial relationship of local regions. The efficiency is preserved by the separation of individual words in a visual phrase and the sparsity of the representation. Multiple descriptors of a local region can be assigned to a visual phrase by assigning every descriptor to a visual word separately. The maximum number of nonzero bins in the histogram of visual phrase frequency is the number of local regions such that the representation of image is sparse and the similarity distance is fast to compute. To evaluate the proposed method, we test the proposed method on several benchmark datasets [1], [9] built from TRECVID 2003–2006 corpus [10]. To the best of our knowledge, the proposed model is not only most accurate but also efficient compared to the "bag-of-word" model and the state-of-the-art algorithms [2], [11], [12] on all datasets. We also conduct the evaluation of hierarchical version of CPM on UKBench dataset in [13]. Using the pyramid matching kernel (PMK) [14], [15], the proposed hierarchical CPM achieves significant improvement over hierarchical BoW model by enforcing local coherence properties.

The rest of the paper is organized as follows. We review the related work about IND retrieval and some representative work of object categorization which is an extension of the BoW model in Section II. Section III gives the detail description of the proposed coherent phrase model including the design of two types of phrase. The whole framework of IND retrieval is summarized in Section IV, including a complexity analysis of the proposed method. The experiment results on benchmark datasets are reported and compared with other methods in Section V. Finally, we conclude with the contributions of this paper in Section VI.

II. Related Work

Although previous research about exact duplicate and copy detection are mainly based on image representation using global feature, e.g., color, some researchers have proposed to use BoW model of images to detect and retrieve INDs. Zhang and Shi [1] proposed to identify IND using a stochastic attributed relational graph (ARG) matching. They learnt a distribution-based similarity from the spatial relation among local interest points for matching ARGs. However, the matching speed of this method is slow and the parameters can only be tuned heuristically. Different from [1], Ke *et al.* [16] represented each image as a set of local covariant regions, each of which is characterized by a *PCA-SIFT* descriptor. Locality sensitive hashing (LSH) is used to design an efficient index structure for fact point set matching. The authors employed RANSAC to perform the geometry verification. The robustness of this technique cannot be guaranteed for partial matching due to the cluttered background if geometry verification, which itself is a computationally expensive process, is not carried out. Compared to standard BoW, RANSAC is computationally inefficient. Recently, Zhao *et al.* [2] extended the matching of local point set by introducing one-to-one symmetric property for matching and LIP-IS index structure for fast approximate search. Although it is interesting to learn IND pattern from the histogram of matching orientation, the methods proposed in [2] and [3] are only partially invariant to rotation/scaling transformation and require explicitly calculating point-to-point matching when comparing two images.

On the other hand, BoW model [17] is a promising framework for IND retrieval as well as for generic image categorization. Within this framework, a collection of local regions are either detected [13] or sampled from every image, each of which is represented by a local descriptor, e.g., SIFT [5]. A visual vocabulary is then built by clustering all the descriptors into K clusters, each of which corresponds to a visual word. Finally, each image is represented as a histogram of visual word frequency after assigning every local descriptor to some visual word. The most critical problem of such methods is that the ambiguous visual words will introduce large number of false matches when each region is matched independently of others. Several methods have been proposed to improve this model by capturing the spatial arrangement of visual words. Lazebnik *et al.* [11] extended the pyramid matching kernel (PMK) [14], [15] by incorporating the spatial information of local regions. Two images are partitioned into increasingly fine subregions and PMK is used to compare corresponding subregions. This method implicitly assumes the correspondences between subregions, which is not translation/scale/rotation invariant. Also by utilizing spatial information of local regions, Savarese *et al.* [18] used *correlagram* to measure the distribution of the proximities between all pairs of visual words and used for category classification. But this method is sensitive to spatial deformation. More robustly, the proximity distribution kernels (PDK) proposed in [19] capture only the rank information of the proximities between two visual words. But the rank about proximity is not discriminant and easy to be confused. These methods do not explicitly reduce false matches when assigning local region to visual words. Instead, they improve BoW model by utilizing spatial information and geometric relationship of local regions for matching images. Another way to improve BoW model is to capture the spatial co-occurrence pattern(s) of multiple visual words. Sivic *et al.* [20] used neighborhood structure of local

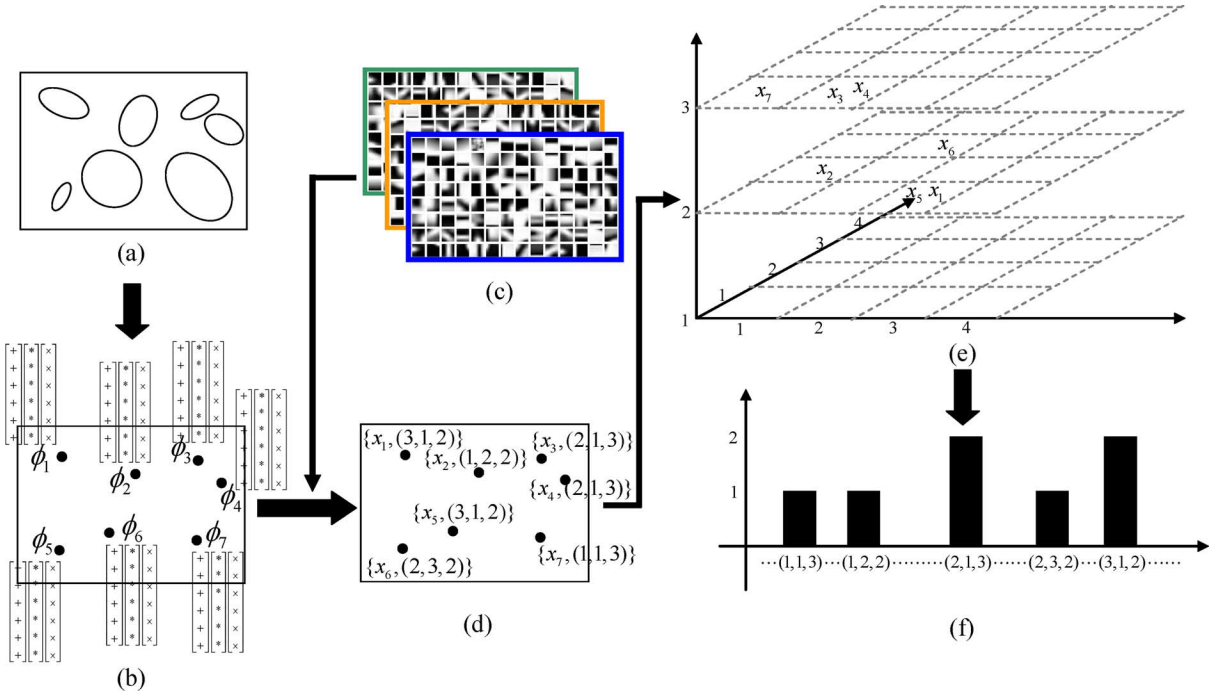


Fig. 2. Demonstration of generating the coherent phrase model of order 3 for an image. (a) Input image I where a set of local regions are located. (b) $K = 3$ descriptors ϕ_i are extracted for every local region i. (c) Three codebooks are built by applying k-means clustering on every type of descriptors. (d)-(e) Every local region is assigned to a visual phrase (3-tuple) according to the codebooks. (f) Final representation of image I, which is a sparse histogram of visual phrase frequency. (a) Image and local regions. (b) Multiple descriptors of local regions. (c) $\mathrm{K} = 3$ codebooks of visual word. (d) Association of interesting point with K-tuple of visual phrase. (e) Assignment of interesting points to visual phrase. (f) Histogram of visual phrase frequency.

points as a neighboring consensus constraint for retrieving duplicate objects from video. Similar to the proposed SCP, spatial coherency is enforced for duplicate object detection in this method. However, [20] analyzed the spatial coherency among multiple regions while the proposed SCP examines the spatial coherency within single regions leading to a computationally efficient algorithm. References [21] and [22] applied the frequent itemset mining to discover meaningful patterns which are discriminant for object classification. The common disadvantage of [18], [19], [21], and [22] is that the number of combinations of multiple words is exponential to the number of local regions, which is not optimally efficient. Note that our proposed method is to describe every local region using visual phrase (multiple descriptors) instead of visual word (single descriptor) and remove the false matches of local regions by the coherence property when assigning visual region to visual phrase. All the above methods can be combined with our method to further improve their performance.

For the problem of IND retrieval, [2] and [12] have achieved the promising performance. The standard "bag-of-word" model and its extensions, e.g., the fusion method [23] combining SIFT and SPIN descriptors and SPM [11], are widely used in object categorization. We will compare our method with these relevant methods in the experiment section.

III. Coherent Phrase Model

Standard BoW model, although efficient, ignores the spatial information [18] and results in the ambiguity of visual words [21]. In this section, we propose a new coherent phrase model to enhance the standard BoW model. Given an image, M local regions can be located by using some detector [24] or using dense sampling [11]. For every local region i, a combination of K descriptors ϕ_{ik} are extracted from the local region to characterize its appearances:

$$\phi_i = \{\phi_{i1}, \phi_{i2}, \ldots, \phi_{iK}\}. \tag{1}$$

Every $\phi_{ik}, k \in [1, 2, \ldots, K]$ belongs to one of the visual words and each ϕ_i is mapped to a K-tuple of visual words. We call this K-tuple as a **visual phrase** v

$$v = \{v_k | k \in [1, 2, \ldots, K]\} \tag{2}$$

where each v_k indicates a visual word. K is denoted as the order of visual phrase. By assigning every ϕ_i to the corresponding visual phrase, we obtain the coherent phrase model of an image, which is a histogram of visual phrase frequency. Fig. 2 illustrates the process of generating this representation where the order of visual phrase is 3.

The distinct property of the coherent phrase model is that under this model, every local region is characterized by *multiple* descriptors. Multiple descriptors of every local region describe different aspects of the appearance. Hence, we can incorporate the meaningful coherency across multiple descriptors of every local region. Specifically, two local regions are recognized as

a matched pair only when all of K descriptors are matched to each other:

$$\phi_i \doteq \phi_j \Longleftrightarrow \bigcap_{k=1}^{K}(\phi_{ik} \doteq \phi_{jk}),\ \forall\, i,j \in [1,\ldots,M] \quad (3)$$

where $\doteq$ denotes the match indicator. Since every pair of descriptors (ϕ_{ik} and ϕ_{jk}) match to each other, they belong to the same visual word v_k and ϕ_i and ϕ_j belong to the same visual phrase v. Thus, we can find the matching of two local regions efficiently by assigning them to visual phrase. Different types of coherency can be enforced by designing different types of visual phrase, where K descriptors are generated in different ways. In the following subsections, we propose two types of visual phrase, where feature and spatial coherencies can be enforced, respectively.

A. Feature Coherent Phrase

Multiple types of local descriptors [5], [25], [26] have been proposed in the literature to represent local image region. According to the study in [27], there is no single descriptor which is superior to others because different descriptors represent different features about the local region. For example, SIFT [5] and its variants, e.g., [28], record the statistics of gradient orientation, SPIN [26] encodes the distribution of intensity at different distances from a center, and Shape Context [25] describes the information of shape.

If we extract multiple types of local descriptors from a local region, we can construct the visual phrase consisting of different types of descriptors, which is called **feature coherent phrase (FCP)**. For example, we can generate FCP of order 2 as follows:

$$\phi_i(2) = \{\phi_{i1} = D_{sift}(R_i), \phi_{i2} = D_{spin}(R_i)\} \quad (4)$$

where R_i indicates the i^{th} local region, $D_{sift}(R_i)$ and $D_{spin}(R_i)$ are the SIFT and SPIN descriptors extracted from R_i, respectively. FCP enforces the coherency across different types of feature between two matched local region. When matching two FCPs using (3), it is required that every feature of two local regions needs to be matched for matching these two regions. Note that FCP is not limited by the descriptor it used, which will be verified empirically in the experiment section.

For descriptors of kth type used in FCP, we perform k-means clustering to obtain V_k visual words of kth type $\mathcal{V}_k = \{v_1, v_2, \ldots, v_{V_k}\}$. K codebooks of visual word are obtained for K different types of descriptor. Notice that the codebook size V_k can be different for different types of descriptor. For every local region, we map K descriptors to an FCP (K-tuple of visual words) by assigning every descriptor to some visual word in the corresponding codebook of same type separately.

B. Spatial Coherent Phrase

Every local region is associated with a scale, which is either provided by the detector or fixed in dense sampling case. Local descriptor is commonly extracted from a region of such scale. We denote this scale as *description scale*. According to the assumption of spatial coherency [29], the neighborhoods of two matched regions are likely to match to each other. This motivates the design of **spatial coherent phrase (SCP)**.

Given a local region i as well as the associated description scale s_0, we expand s_0 to multiple expanded scales as

$$s_k = C \cdot s_{k-1},\ k \in [1,2,\ldots,K] \quad (5)$$

where the constant C controls the expansion ratio between two consecutive scales to include the neighborhood of local regions. We set the value of C close to 1 and only use the additional scales which are near to the original description scale to emphasize local spatial coherency for reducing the effect of irrelevant background. K descriptors are extracted from the expanded regions of these K scales using only one type of descriptor

$$\phi_i = \{\phi_{ik} = D(R_{i,s_k}) | k \in [1,2,\ldots,K]\} \quad (6)$$

where R_{i,s_k} denotes the local region of scale s_k centered at the same center of the local region i and $D(R_{i,s_k})$ denotes the descriptor (e.g., SIFT) extracted from R_{i,s_k}. SCP enforces the coherency across different spatial neighborhoods of two matched regions. Two local regions are matched only when their spatial neighborhoods of K expanded scales centered at the same center are consistently matched.

When selecting the order (K) of SCP, there are two factors to be considered. Given the codebook size of single visual word, the order K of SCP controls the codebook size of visual phrase. A suitable K needs to be selected to generate desired number of visual phrase. Given the expansion ratio C, the order K of SCP controls the spatial context of every local region. Because spatial coherency is only applicable in the local context, a suitable K needs to be selected to constrain the size of local context. Because the codebook size of each dimension is 500 in our experiment, we use only $K = 2$, which generates a codebook of $500^2 = 250\,000$ visual phrases. By reducing the codebook size of each dimension, e.g., from 500 to 50, we can design SCP of higher order. We will conduct an experiment to analyze how the order K of SCP affects the IND retrieval performance in the experiment section.

Note that SCP strictly utilizes the description scale from the detector and expands it accordingly using a fixed expansion ratio. SCP preserves the chance of two local regions to get matched as the same as from the detector. This is because the original description scale output from the detector is scale invariant, which allows for repeatedly detecting two matched regions in different scales. By expanding those two description scales using a fixed expansion factor, two corresponding regions in the new expansion scales are the matched region, and hence, their descriptors in two corresponding expansion scales can be matched. SCP uses the characteristic scales of the regions output from detector as the reference for expanding scales using a fixed expansion factor. It achieves the same invariant ability to scale transformation as the detector. Fig. 3 illustrates this property of SCP using two pairs of images. We can see that the same regions appear at different scales

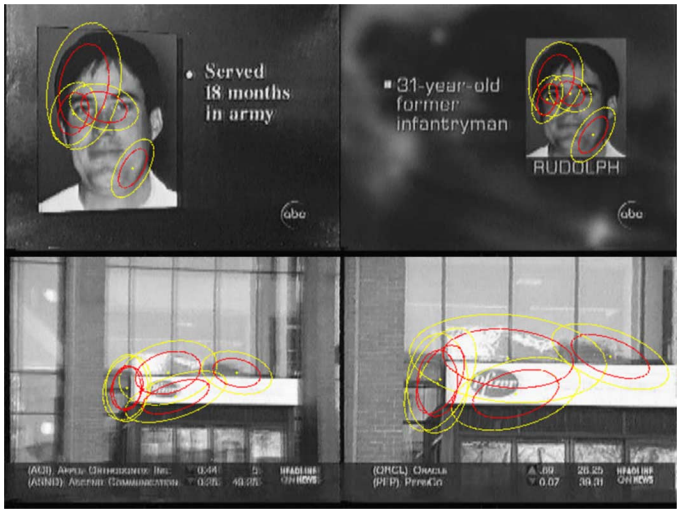
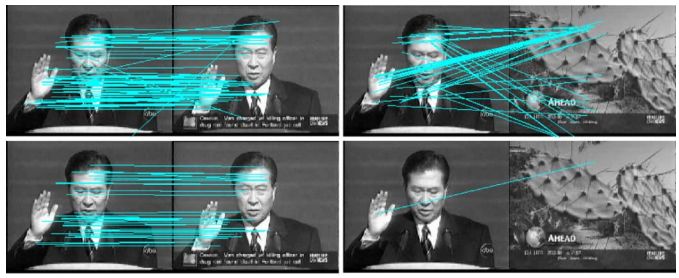


Fig. 3. Two pairs of example show that SCP preserves the scale-invariant property of local region detector. The ellipses in red are the local regions output from detector in description scale s_0, and the ellipses in yellow are the regions in expanded scales $s_1 = 1.2 * s_0$ from SCP. The corresponding regions are matched in both description scale and expanded scale.

in two images because they undergo a scale transformation. The detector output these two scales (red) as the characteristic scales of them, which allow for detecting the matched region in two images. By expanding the two characteristic scales using a fixed expansion factor (e.g., 1.2), we obtain two matched regions in corresponding expanded scales (yellow) because the expansion uses the characteristic scales from detected region as the reference. At every corresponding expanded scale, the obtained regions can be matched in the same way. Hence, SCP preserves the scale-invariant property of the detector and allows for matching the regions undergoing different transformation. Similar to FCP, we perform k-means clustering on all the available descriptors of one type used in SCP. The obtained V_k visual words $\mathcal{V}_k = \{v_1, v_2, \ldots, v_{V_k}\}$ form a codebook of single type of descriptor. A local region is mapped to an SCP (K-tuple of visual words of the same type) by assigning K descriptors to some visual word in the single codebook separately.

The proposed coherent phrase model provides a general framework to incorporate coherency across multiple descriptors of every local region. The proposed FCP introduces the feature coherency across different types of descriptors, and SCP introduces the spatial coherency across neighborhoods of different scales. Both of these coherency can effectively reduce the number of false matches due to the ambiguity of single local descriptor and the errors caused by k-means clustering. Fig. 4 shows an example of matching using visual word and SCP of order 2. We can see that the numbers of matches in both pairs of images are close when visual word of single descriptor is used. However, the number of matches between the IND pair is much more than that between the pair which is not IND when we use the proposed SCP. Many false matches between the pair of images that are not IND pair are removed because of the increased discriminant power of the proposed coherent phrase model. Besides FCP and SCP, it is possible to design other visual phrase to incorporate other forms of coherency. One example is to combine the proposed FCP and SCP into a hybrid phrase which considers both feature and spatial coherency.



(a) (b)

Fig. 4. Example of matching between (a) matched images and (b) unmatched images. The first row are the results using visual word of single SIFT descriptor, and the second row are the results using SCP consisting of two SIFT descriptors.

C. Hierarchical CPM

The proposed CPM can be extended to its hierarchical version. Similar to [14] and [15], the whole feature space of single descriptors is partitioned into a hierarchical tree by hierarchical clustering algorithm (e.g., hierarchical K-means). Two parameters control this partition: the number of tree levels L and the branch factor m. In level l, there are $m^{(l-1)}$ clusters. The hierarchical CPM can then be directly derived from this hierarchical tree of visual codebook. Every local region is assigned to a hierarchy of visual phrase by assigning each individual descriptor to the hierarchy of the corresponding type of visual word. In different resolution, the combination of multiple visual words to which the corresponding descriptors belong forms the visual phrase at this level. All the visual phrases in different resolutions form a hierarchical pyramid of visual phrase. By assigning every local region to visual phrase in different levels, we can generate the hierarchical histogram of visual phrase for every image, which is similar to hierarchical histogram of visual word of pyramid matching kernel. The similarity between images can then be measured by any similarity measure such as the pyramid matching kernel [14], [15] using this representation.

IV. Near-Duplicate Image Retrieval

We use a similar framework to [13] and [30] for IND retrieval. It consists of two parts: the process for database initialization and the process for handling input query. The offline process for database initialization includes the following.

1) Locate local regions, and extract descriptors from every local region for all images in the database. Here we use dense sampling [11] to extract descriptors on an overlapped 8×8 grid to cover the whole image. It avoids the risk of losing some information by using only specific detector and also avoids the detection of interest points. In terms of descriptors, SIFT and SPIN descriptors are used for FCP, and only SIFT descriptor is used for SCP.
2) Build the codebook(s) for different types of descriptors. We apply k-means clustering on all descriptors of every type and generate the codebook(s). Note that hierarchical k-means clustering can be used to generate codebook and histogram of multi-resolution.

3) Assign local regions to visual phrases, and form the histogram of visual phrase frequency for every image. For every local region, it is assigned to a visual phrase by assigning every individual descriptor to some visual word of the same type separately and obtaining the K-tuple of visual phrase.

For online query processing, the histogram of visual phrase frequency is calculated in a similar way for the query image. It is then used to calculate the similarity between the query image and every image in the database. The similarity can be calculated by using any distance for two histograms, e.g., $L2$ distance, χ^2 distance as well as EMD distance, etc. In this paper, we use the intersection distance τ to measure the similarity between two histograms H_I and H_J of visual phrase frequency:

$$\tau(H_I, H_J) = \sum_{v=1}^{V} \min(H_v(I), H_v(J)) \tag{7}$$

where $H_v(\cdot)$ represents the vth bin of the histogram. Although simple, the intersection distance can handle partial matching with cluttered background. It can be replaced by PMK [14] if we extend to multi-resolution histogram by considering hierarchical codebook(s). In this paper, we only use the simple intersection distance to show the effectiveness of the proposed coherent phrase model.

A. Complexity Analysis

Compared to the retrieval frameworks [13], [30] which use the standard BoW model, the proposed method preserves the efficiency, especially for test query comparison. Although we explore the space of visual phrase, which is a combination of K visual words, there is no need to build the codebook of visual phrase. Because we *separately* assign individual descriptor to visual word, only the codebook(s) of visual word are built, whose size is much smaller than the actual number of all possible visual phrase. For example, we only build two codebooks with the same size of 500 for SIFT/SPIN, respectively, and local regions can be mapped to $500 \times 500 = 250\,000$ different visual phrases of order 2. Compared to the standard BoW model, the performance of the proposed model does not degrade heavily using smaller codebooks of visual word. It is because visual phrase combines K visual words and has larger discriminant power. This advantage can reduce the computation complexity of both offline quantization and assigning local regions to visual phrases. When assigning individual descriptor of a local region to visual word, the number of required comparison is reduced by using small codebook. Because the process to initialize database can be completed before the query is submitted, its complexity does not affect the efficiency for online process of query.

For test query comparison, the proposed image representation using visual phrase preserves the sparsity of BoW representation. Suppose there are M local regions in an image, the maximum number of nonzero bins in its histogram of visual phrase frequency is M, which is the same as that in BoW representation. Because of this sparsity, the computation of the similarity between two images is very efficient. Notice that graph-based method like [2] requires us to explicitly find the correspondences between two images and will result in a higher computational complexity. For the methods that utilize the configurations between K local regions [18], [20], [21], the maximum nonzero bins of the histogram in their methods are exponentially increased with M (i.e., M^K). For example, [19] that used the rank information about the proximity between two regions has the maximum M^2 of nonzero bins in the histogram. Compared to our proposed method, these methods are much more computationally expensive.

The only bottleneck of the proposed method is about the additional cost for the extraction of multiple features. In the initialization of database, we need to extract and assign multiple descriptors instead of single descriptors for every local region in step 1 and 3. The time cost for feature extraction increases *linearly* when more types of feature descriptors are involved. Because most of the feature extraction is done during the initialization of database, which is only performed once in training phrase, such additional cost for feature extraction does not heavily affect the efficiency of our proposed method. By using distributed computing environment, multiple features can be computed in a parallel fashion, which can further eliminate such additional cost.

V. Experiment Evaluation

A. Experiment Setup

We evaluate the proposed coherent phrase model of FCP and SCP for IND retrieval. Four datasets are used to compare the proposed method with the latest methods reported in the literature. Three of them are the IND datasets which are extracted from the keyframes of the TRECVID corpus [10]. Two small datasets are used for detailed analysis: Columbia's TRECVID2003 dataset provided by [1] and NTU's TRECVID 2005&2006 dataset, which contains the keyframes extracted from TRECVID 2005 & 2006 corpus where we introduce large view angle and scale changes. Both datasets include 150 IND pairs (300 images) and 300 nonduplicate images. The size of all the images are 352×240. We also evaluate the proposed method on the larger CityU's TRECVID2004 dataset [9] which contains 7006 keyframes from TRECVID 2004 corpus. The fourth dataset is the UKBench dataset [13] which contains a total of 10 200 images about 2550 objects. Both of two common feature extraction strategies for BoW are applied in the experiments: 1) For TRECVID datasets, 8×8 local regions extracted from dense sampling with spacing of 4 pixels are used to generated FCP/SCP; the number of local regions detected in one image is 5280 for FCP and $5280 \times 2 = 10\,560$ for SCP of order 2; 2) we used the local regions extracted by using *SURF* detector to generate FCP/SCP for UKBench dataset. The number of local regions detected in one image varies in the range from 40 and 2000. It is shown that the proposed CPM generally improves the IND retrieval performance, independent of different feature extraction methods. For FCP, we use SIFT and SPIN descriptors to extract FCP of order 2. For SCP, we set $C = 2$ and use two scales (s_0 and $s_1 = 2 \cdot s_0$) to extract SCP of order 2 using SIFT/SURF descriptor. The sizes of all codebooks are fixed as 500 in FCP, SCP, and in all the other methods used

for comparison. To compare with the state-of-the-art, we adopt the same evaluation protocol used in [2] and [12]. All IND pairs are used as queries to evaluate performance. For each query, we calculate the similarity between the query image and all images using the intersection between their histograms of visual phrase frequency. A ranked list of images is then produced according to their similarities to the query. To evaluate the retrieval performance and compare with other methods, we estimate the probability of the successful retrieval $P(n)$ by averaging cumulative top-n accuracy as

$$P(n) = \frac{Q_c}{Q} \tag{8}$$

where Q_c is the number of queries that rank their INDs within the top-n position, and Q is the total number of queries. Note that $P(n)$ is equal to the precision/recall value at the knee point (where precision is equal to recall) if there are n IND correspondences for every query image.

B. Evaluation on Small Datasets

We perform solid evaluation for the proposed CPM on two small datasets: Columbia's TRECVID2004 dataset and NTU's TRECVID2005&2006 dataset. We compare FCP with several methods: 1) the state-of-the-art algorithm [2] which used one-to-one graph matching; 2) the standard BoW model using individual SIFT and SPIN descriptors, denoted as SIFT BoW and SPIN BoW. To prove the advantage of incorporating feature coherency using FCP, we also compare FCP with the fusion method used in [17] to combine the two BoW models using SIFT and SPIN. The fusion is computed as

$$S_f(I,J) = \sum_{l=1}^{K} \frac{\alpha_l}{1+\exp\left(-S_l(I,J)\right)} \tag{9}$$

where $S_f(I,J)$ denotes the similarity between image I and J after fusion, and $S_l(I,J)$ denotes the original similarity calculated in each BoW model using individual descriptor, e.g., SIFT/SPIN. In our experiment, the parameters are set as $K=2$ and $\alpha_1 = \alpha_2 = 0.5$. For SCP, we compare it with two methods: 1) the standard BoW model using SIFT descriptor, denoted as SIFT BoW; 2) the spatial pyramid matching (SPM) method [11] which also utilizes spatial information. We do not compare SCP with the fusion method since SCP only uses one type of descriptor, i.e., SIFT. Actually, our SCP outperforms the fusion method combining SIFT and SPIN descriptors.

Fig. 5 shows the results of these methods for top-n IND retrieval on Columbia's TRECVID2004 dataset where n changes from 1 to 20. In terms of FCP, we can see that SIFT descriptor achieves better performance than SPIN descriptor when using BoW model. Although the fusion of two BoW models can improve the overall performance, the performance after fusion is still lower than the result of [2]. However, our proposed FCP using SIFT and SPIN achieves a significant improvement (8% for top-1) over the fusion method and outperforms [2] by 4.3% for top-1 retrieval. One issue about the Columbia's TRECVID2003 dataset is that 5% of the annotated IND pairs in

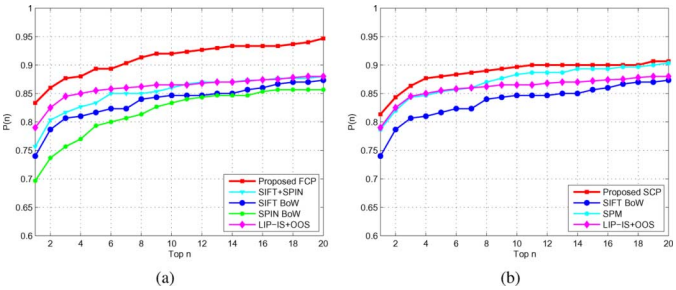


Fig. 5. (a) Top-n IND retrieval performance of FCP and relevant methods on the Columbia's TRECVID2003 dataset. (b) Top-n IND retrieval performance of SCP and relevant methods on the Columbia's TRECVID2003 dataset.

Fig. 6. Four labeled IND pairs (first two rows) from Columbia's TRECVID2003 dataset and the most similar image to the images in the first row returned by the proposed FCP.

the provided ground-truth annotations are confusing according to our observation. Fig. 6 shows four examples where the images returned by the proposed method are different from the annotated INDs. However, the returned images are also visually correct INDs in our view. The reported performance of the proposed method is strictly based on the provided ground-truth annotations. However, the ground-truth annotation itself contains some errors. If these errors are removed, the accuracy of retrieval using the proposed method will be increased further.

In terms of SCP, although SPM outperforms the standard BoW model using SIFT descriptor, the proposed SCP achieves even higher performances. This is because that the mechanisms of using spatial information in SCP and SPM are different. SPM constrains the matching in the corresponding spatial grids, which is not invariant to object translation/rotation and the changes of viewpoint and scale. Similar performances are also achieved on NTU's TRECVID2005&2006 dataset. Fig. 7 shows four IND pairs from this dataset where there exists rotation and large changes of scale and viewpoint. SCP outperforms SPM in these cases because it is invariant to these changes while the spatial constraints in SPM are wrong in the presence of these changes. Note that in Fig. 7, some IND examples (first and second columns) contain substantial viewpoint changes and some examples (first, second and third columns) contain large scale transformation between IND pair such that only the sub-image(s) form the IND counterpart. While the scale of two images in the first, second, and fourth columns are extremely different, SCP still can correctly detect them as the IND pairs. It is empirically shown that the proposed SCP works well with substantial scale/viewpoint changes and is capable to discover the correct INDs which may only contain sub-image(s). We

TABLE I
TOP-1 IND RETRIEVAL PERFORMANCES OF DIFFERENT METHODS ON COLUMBIA'S TRECVID2003 DATASET (COLUMBIA) AND NTU'S TRECVID2005&2006 DATASET (NTU)

Methods	**Proposed FCP**	**Proposed SCP**	SIFT BoW	SPIN BoW	SIFT+SPIN Fusion	LIP-IS+OOS [2]	SPM [11]
Columbia	**0.833**	**0.813**	0.740	0.697	0.756	0.790	0.787
NTU	**0.960**	**0.960**	0.920	0.843	0.937	/	0.940

Fig. 7. Four labeled IND pairs (first two rows) from NTU's TRECVID2005&2006 dataset where SCP succeeds to return them as top-1 images but SPM fails. The top-1 images returned by SPM are shown in the third row.

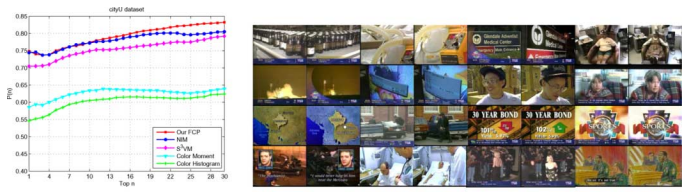


Fig. 8. (a) Top-n IND retrieval performance of FCP and relevant methods on the CityU's TRECVID2004 dataset. (b) Examples of IND pairs returned by the proposed FCP where scale, illumination, viewpoint, as well as image editing changes occur.

also summarize the top-1 IND retrieval performances of all relevant methods on both datasets in Table I. Because every query image only has one IND correspondence in these two datasets, the top-1 retrieval accuracy reported in Table I is equal to precision/recall value at the knee point. The results are consistent across the two datasets. The performance of the fusion method that combines two BoW models using SIFT and SPIN is only comparable to the best performance of individual model. The LIP-IS+OOS method [2] and SPM [11] are two state-of-the-art methods which achieve better performances than the fusion method. But our proposed models of FCP and SCP significantly outperform any of the existing methods on both two datasets.

C. Evaluation on Large Datasets

To illustrate the general validity of the proposed method, we also evaluate CPM on two large datasets: CityU's TRECVID2004 dataset [9] (7006 images) and UKBench dataset [13] (10 200 images).

1) IND Retrieval on CityU's Trecvid2004 Dataset: We applied FCP on the CityU's TRECVID2004 dataset for IND retrieval. Different from two small datasets from TRECVID corpus, images have multiple IND correspondences. We extracted both SIFT and SPIN descriptors from densely sampled regions in multiple scales and generated FCP of order 2. We compare our FCP with several methods which have been reported on this dataset: 1) some conventional approaches using global features, e.g., color moment [31] and color histogram [1]; 2) nonrigid image matching (NIM) [12] which is the latest state-of-the-art on CityU's TRECVID2004 dataset. Fig. 8 shows the results of top-n IND retrieval on CityU's TRECVID2004 dataset where n changes from 1 to 30. We can see that for the small n, the proposed FCP achieves the very similar accuracy with NIM and it outperforms NIM when n becomes large. We also show the precision/recall values at knee point for different methods in Table II. To the best of our knowledge, FCP achieves the best performance on CityU's TRECVID2004 dataset. Note that the results reported for NIM are based on multiple features including both global features as well as local descriptors, and this approach requires explicit image matching. The proposed method only uses local descriptors and does not calculate image matching, which leads to a more efficient solution. Here, LIP-IS+OOS in [2] is not compared because it is computationally too expensive to be applied on this dataset.

2) Object Retrieval on Ukbench: We also test the proposed method on UKBench dataset [13] in the context of object retrieval which is an example of IND retrieval. This dataset contain 10 200 images where every four images capture a single object with different viewpoint, illumination condition, and scale. There is a total of 2550 different objects such as CD covers, etc. The size of all images are 640×480. We locate local regions and generate the corresponding descriptors for every image in the dataset using *SURF* [32]. The purpose of using this set of new detector/descriptor is to illustrate that the proposed method can be generally applied on any local representation. The key to the performance improvement is the coherence property of the visual phrase scheme and not the power of descriptor. In this experiment, we evaluate hierarchical SCP (H-SCP) of order 2 and use the pyramid matching kernel [14], [15] as similarity distance to compare two images.

The experiment setup for object retrieval is similar to the IND retrieval mentioned before. We setup the databases of different size by sequentially selecting images from UKBench dataset. The smallest database contains the first 1000 images of this dataset while the largest database contains all the images. For every database, we generate a hierarchical vocabulary tree ($L = 6$ and $m = 10$) where there are 10^5 leaf nodes. Different from the performance reported in [13], we use all the images in the database as query to evaluate the performance, which can avoid the effect of query selection and better illustrate the general

TABLE II
PRECISION/RECALL VALUES AT THE KNEE POINT (WHERE PRECISION IS EQUAL TO RECALL) ON CITYU'S TRECVID2004 DATASET **AUTHOR: PLEASE CITE TABLE IN BODY OF PAPER.**

FCP	NIM [12]	S^3VM [12]	Color Moment [31]	Color Histogram [1]
0.7489	0.7442	0.7049	0.5864	0.5463

TABLE III
TOP-4 RECALL OF HIERARCHICAL SCP (H-SCP) AND HIERARCHICAL BAG-OF-WORDS (H-BOW) USING SURF ON UKBENCH DATASET

	1000 images	3000 images	5000 images	7000 images	10200 images
H-SCP+PMK	**0.7458**	**0.7302**	**0.7520**	**0.7494**	**0.7503**
H-BoW+PMK	0.5912	0.5691	0.5831	0.6173	0.6155

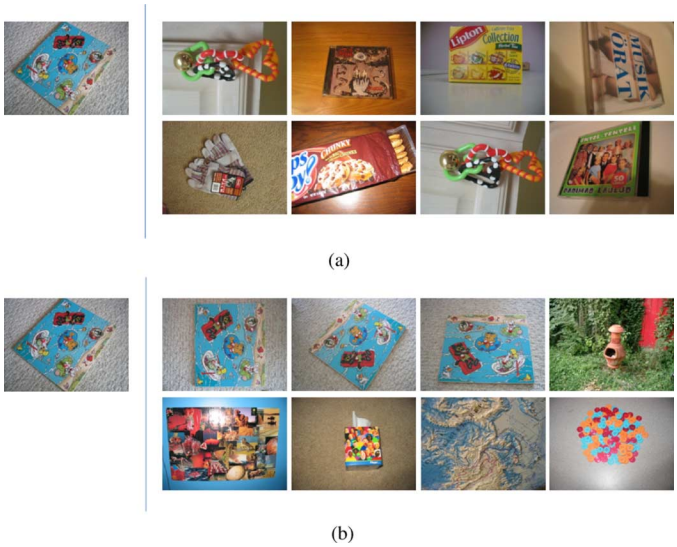


Fig. 9. Example queries and top-8 similar images returned by (a) hierarchical BoW and (b) hierarchical SCP under the same configuration (database of 10 200 images represented using SURF, a vocabulary tree of six levels with a branch factor 10).

performance. Similar to the setup for IND retrieval, we compare every images in the database with the query and obtain a rank list. For every query, there are totally four images (including the query itself) capturing the same object. To evaluate the performance of retrieving multiple IND, we modify the measure in [13] by normalizing it with four (the total number of ground truth INDs for a query) for the top-4 retrieval. This measure counts the percentage of the four ground truths of a query in the top-4 retrieved images, which is also the precision/recall value at the knee point. Table III compares the top-4 retrieval performances of the proposed H-SCP and the hierarchical bag-of-words (H-BoW) representation of SURF using the same PMK. It is shown that H-SCP consistently outperforms H-BoW representation across different scales of database. Compared with H-BoW, more INDs are returned within top-4 images of H-SCP. The average improvement of the proposed method is 15.6%, which is quite significant. Note that both methods use the *same* detector/descriptor, the same hierarchical vocabulary tree, and the same matching kernel to make the comparison fair. This fact also proves that the improvement is indeed caused not by other factors, but by the coherence property enforced by the proposed phrase structure. From the example shown in Fig. 9, we can see that the proposed method not only retrieves all the four images containing the same object but also finds other similar images which the BoW representation cannot achieve. Note that we are unable to directly compare with [13] because the used detector/descriptor and the queries used for performance evaluation are not mentioned in the paper, which do not allow for direct comparison.

D. Complexity Comparison

According to the analysis of Section IV-A, the coherent phrase model has a low computational complexity for calculating the similarity between two images. We evaluate the average CPU time for calculating the similarity between two images for different methods. For SPM and FCP/SCP, we estimate the average time of calculating similarities for Columbia's TRECVID2003 dataset using our own matlab implementation without optimization. For LIP-IS+OOS method in [2], we use the C++ implementation provided by the authors to estimate the average time. On the UKBench dataset, we also evaluate the average CPU time for calculating pairwise similarity of H-SCP and the original H-BoW based on the C++ implementation of [33]. From the average time reported in Table IV, it is empirically shown that the coherent phrase models of FCP and SCP are not only more accurate but also more efficient than other methods. Notice that the LIP-IS+OOS method in [2] is least efficient because it requires to find the exact correspondence for every local region when calculating the similarity between two images. The average time of SPM is longer than FCP/SCP using the codebook of same size. It is because SPM needs to compare multiple pairs of histograms of different subregions at different levels while FCP and SCP only compare one pair of histograms about the whole image. In terms of H-SCP, it is shown that the proposed method preserves the efficiency of standard BoW model, and the C++ implementation allows comparing two images very efficiently. We admit that there is an additional overload due to the increase in the number of features. However, most of such overload in feature extraction and assignment is done offline for database and can be neglected.

Another factor that will affect the complexity of IND retrieval is the codebook size. Fewer visual words need to be compared when every local region is assigned to a visual phrase/word of a smaller codebook. But the visual words of a smaller codebook will introduce more ambiguities and hence degrade the performance. We evaluate the performance degradation of our

TABLE IV
AVERAGE CPU TIME FOR CALCULATING SIMILARITY BETWEEN TWO IMAGES

	Dense Grid on TRECVID (Matlab)			Sparse Region on UKBench (C++)	
	FCP/SCP	LIP-IS+OOS [2]	SPM [11]	H-SCP	H-BoW
Time(ms)	**16.1**	130	24.3	**0.09**	0.08

TABLE V
PERFORMANCE DEGRADATION OF TOP-1 IND RETRIEVAL WHEN THE CODEBOOK SIZE IS REDUCED

	From 500 to 100 (TRECVID)			Level from 6 to 3 (UKBench)	
	FCP/SCP	SIFT BoW	SIFT+SPIN Fusion	H-SCP	H-BoW
Degrade	**4.0%**	17.7%	7.1%	**8.5%**	15.2%

TABLE VI
TOP-1 IND RETRIEVAL PERFORMANCES OF SCP OF DIFFERENT ORDERS ON COLUMBIA'S TRECVID2003 DATASET

SCP Order	K=1	K=2	K=3	K=4	K=5
Top-1 Accuracy	0.687	0.733	**0.817**	0.770	0.757

proposed FCP/SCP on both Columbia's TRECVID2003 dataset when the codebook size is reduced from 500 to 100. From the results shown in Table V, we can see that the performance degradation of FCP/SCP is much smaller than BoW model using SIFT and the fusion method combining SIFT and SPIN. On UKBench dataset, we similarly reduce the hierarchical codebook from six levels to three levels while keeping the branch factor as 10. The number of leaf nodes (finest visual words) is reduced from 10^5 to 10^2, respectively. We can also see from Table V that H-SCP is much more robust than H-BoW model when the smaller codebook is used. This experiment empirically show that the proposed FCP/SCP as well as its hierarchical version H-SCP can improve efficiency further by using smaller codebooks while still maintaining the high retrieval accuracy.

To illustrate how the performance of SCP is affected by its order K, we conduct an experiment to test the performance of SCP from order 2 to 4 and compare with standard BoW (i.e., SCP of order 1). Each dimension of the codebook is fixed as 50 and the expansion ratio C is set as 1.2. Table VI shows the top-1 average precisions of SCP of different orders on Columbia's TRECVID2003 dataset using SIFT as the descriptor. From the results, we can see that SCP(s) of different orders consistently outperform BoW while the best performance is achieved by the SCP of order 3 corresponding to a codebook of total $50^3 = 125\,000$ visual phrases. The reason that SCP(s) of the order higher than 3 degrade the performance is because the codebook size of visual phrase is too large as well as the expanded scales at $k = 4$ and $k = 5$ are too large with respect to the neighborhood. Note that we achieve the same highest performance as in Table I by using the SCP of order 3, where a smaller codebook of $50^3 = 125\,000$ visual phrases is used. This is another factor to further reduce the computational complexity of SCP for IND retrieval.

To verify the fact that the performance of FCP is not limited by the used descriptors, we evaluate the performances of IND retrieval using FCP with different combinations of descriptors. In Fig. 10(a), the combination of a superior descriptor (e.g., SIFT) and a inferior descriptor (e.g., complex filter (CF)—the second worst performance reported in [27]) for FCP is evaluated. In Fig. 10(b), the combination of two inferior descriptors (e.g., cross correlation (CC) and CF—two worst performances in [27]) for FCP is evaluated on Columbia's TRECVID2003 dataset. In both cases, FCP has achieved much higher accuracy for IND retrieval than BoW model using either individual descriptor as well as the fusion method [17] combining two descriptors. This experiment empirically shows that the proposed FCP is not limited by the weak descriptor it used and that the proposed coherent phrase is superior than the traditional visual word.

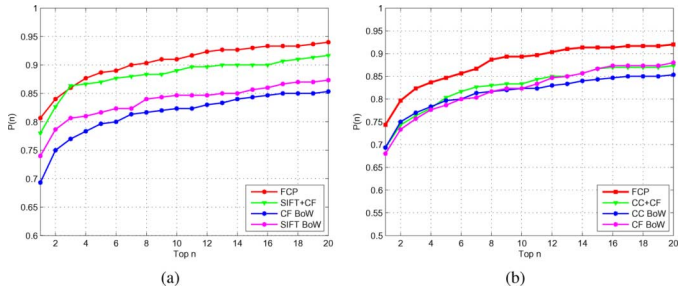


Fig. 10. Top-n IND retrieval performance on Columbia's TRECVID2003 dataset of (a) FCP using SIFT and complex filter (CF). (b) FCP using cross correlation (CC) and CF. Their performances are also compared with BoW models using either individual descriptor as well as the fusion method.

VI. CONCLUSION

In this paper, we propose a coherent phrase model for image near-duplicate retrieval. Different from the standard BoW representation, this model represents every local region using multiple descriptors and enforces the coherency across multiple descriptors for every local region. Two types of visual phrase (feature coherent phrase and spatial coherent phrase) are designed to represent feature and spatial coherency. Both coherency are utilized without increasing the computational complexity. This model, although simple, improves the matching accuracy by reducing the number of false matches and preserves the matching efficiency because of the sparsity of the representation. Instead of mining the association of multiple regions which increases the complexity, the quantization error of BoW is reduced by the coherency in the visual phrase of every single region in our method. Both effectiveness and efficiency of the proposed method have been proved on multiple benchmarks built from

TRECVID corpus for IND retrieval as well as a benchmark for object retrieval. The proposed technique can be generally applied on different variants of the BoW model.

REFERENCES

[1] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attribute relational graph matching with learning," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, Oct. 2004, pp. 877–884.

[2] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 1037–1048, Aug. 2007.

[3] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang, "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, Santa Barbara, CA, Oct. 2006, pp. 845–854.

[4] P. Ghosh, E. D. Gelasca, K. Ramakrisnan, and B. Manjunath, "Duplicate Image Detection in Large Scale Databases," in *Platinum Jubilee Volume*. Kolkata, India: K. Indian Statistical Inst., 2007.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] A. Johnson and M. Hebert, "Using SPIN images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, May 1999.

[7] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1294–1309, Jul. 2009.

[8] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classificatio," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, Jun. 2008..

[9] [Online]. Available: http://vireo.cs.cityu.edu.hk/research/ndk/ndk.html.

[10] *TREC Video Retrieval Evaluation (TRECVID).* [Online]. Available: http://wwwnlpir.nist.gov/projects/trecvid/.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, New York, Jun. 2006, vol. 2, pp. 2169–2178.

[12] J. Zhu, S. C. Hoi, M. R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *Proc. 15th Annu. ACM Int. Conf. Multimedia*, Vancouver, BC, Canada, Oct. 2008.

[13] D. Nistër and H. Stewënius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, New York, Jun. 2006, vol. 2, pp. 2161–2168.

[14] K. Grauman and T. Darrell, "The Pyramid Matching Kernel: Discriminative classification with sets of image features," in *Proc. IEEE Int. Conf. Computer Vision*, Beijing, China, Oct. 2005, vol. 2, pp. 1458–1465.

[15] K. Grauman and T. Darrell, "Approximate correspondences in high dimensions," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, Dec. 2006, vol. 19, pp. 505–512.

[16] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, Oct. 2004, pp. 869–876.

[17] D. Xu and S.-F. Chang, "Visual event recognition in news video using kernel methods with multi-level temporal alignment," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.

[18] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlations," in *Pro. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, New York, Jun. 2006, vol. 2, pp. 2033–2040.

[19] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[20] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Washington, DC, Jun. 2004, vol. 1, pp. 488–495.

[21] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.

[22] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[23] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: An In-Depth Study", Tech. Rep., INRIA, 2005.

[24] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Nov. 2007.

[25] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[26] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[27] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[28] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Washington, DC, Jun. 2004, vol. 1, pp. 511–517.

[29] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent object segmentation and classification," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[30] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Computer Vision*, Nice, France, Oct. 2003, vol. 2, pp. 1470–1477.

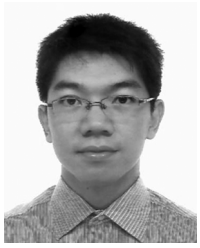
[31] W. Zhao, Y.-G. Jiang, and C.-W. Ngo, "Keyframe retrieval by keypoints: Can point-to-point matching help?," in *Proc. Int. Conf. Image and Video Retrieval*, Tempe, AZ, Jul. 2006, pp. 72–81.

[32] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Computer Vision*, Graz, Austria, May 2006.

[33] J. J. Lee, LIBPMK: A Pyramid Match Toolkit, MIT Computer Science and Artificial Intelligence Laboratory, 2008, Tech. Rep. MIT-CSAIL-TR-2008-17.

Yiqun Hu received the B.S. degree in computer science from Xiamen University, Xiamen, China, in 2002 and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2008.

He is currently the Research Fellow at Nanyang Technological University. His research interests include pattern analysis, machine learning, and their applications in computer vision and multimedia, especially computational attention detection.

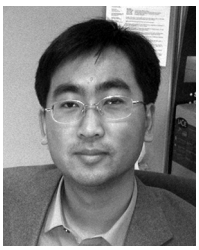


Xiangang Cheng received the B.E. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei. He is currently pursuing the Ph.D. degree in the School of Computer Engineering at Nanyang Technological University.

His research interests include computer vision, multimedia content analysis, machine learning, and pattern recognition.

Liang-Tien Chia (M'90) received the B.S. and Ph.D. degrees from Loughborough University (of Technology), Loughborough, U.K., in 1990 and 1994, respectively.

He is an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. He was the Director, Centre for Multimedia and Network Communications from 2002–2007 and is currently Head, Division of Computer Communications. His research interests can be broadly categorized into the following areas: Internet-related research with emphasis on the semantic web; multimedia understanding through media analysis, annotation, and adaptation; multimodal data fusion; and multimodality ontology for multimedia. He has published over 100 refereed research papers.



Xing Xie (M'04–SM'09) received the B.S. and Ph.D. degrees in computer science from University of Science and Technology of China, Hefei, in 1996 and 2001, respectively.

He is a Lead Researcher in the Web Search and Mining Group of Microsoft Research Asia, Beijing, China, and a guest Ph.D. advisor in the University of Science and Technology of China. He joined Microsoft Research Asia in July 2001, working on spatial data mining, location-based services, and mobile and pervasive computing. He has served on the organizing and program committees of many international conferences such as WWW, GIS, CIKM, MDM, and IUI. During the past years, he has published over 80 referred journal and conference papers.

Deepu Rajan (M'02) received the B.Eng. degree in electronics and communication engineering from Birla Institute of Technology, Ranchi, India, the M.S. degree in electrical engineering from Clemson University, Clemson, SC, and the Ph.D. degree from the Indian Institute of Technology, Bombay, India.

He is an Assistant Professor in the School of Computer Engineering at Nanyang Technological University, Singapore. From 1992 until 2002, he was a Lecturer in the Department of Electronics at Cochin University of Science and Technology, Kochi, India. His research interests include image processing, computer vision, and multimedia signal processing.

Ah-Hwee Tan (SM'04) received the B.S. degree (first class honors) and the M.S. degree in computer and information science from the National University of Singapore and the Ph.D. degree in cognitive and neural systems from Boston University, Boston, MA.

He is currently an Associate Professor and the Head of the Division of Information Systems at the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. Prior to joining NTU, he was a Research Manager at the A*STAR Institute for Infocomm Research (I2R), responsible for the text mining and intelligent agents programs. His current research interests include biologically-inspired cognitive systems, information mining, media fusion, machine learning, and intelligent agents.