Scale-Rotation Invariant Pattern Entropy for Keypoint-Based Near-Duplicate Detection

Wan-Lei Zhao and Chong-Wah Ngo, *Member, IEEE*

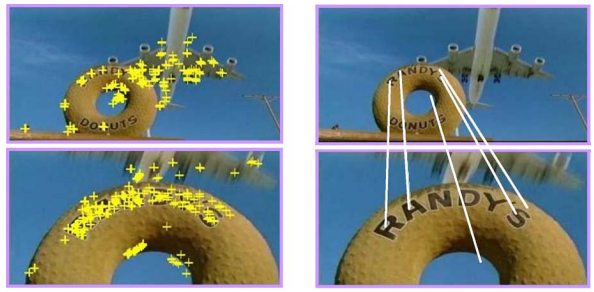
***Abstract*—Near-duplicate (ND) detection appears as a timely issue recently, being regarded as a powerful tool for various emerging applications. In the Web 2.0 environment particularly, the identification of near-duplicates enables the tasks such as copyright enforcement, news topic tracking, image and video search. In this paper, we describe an algorithm, namely Scale-Rotation invariant Pattern Entropy (SR-PE), for the detection of near-duplicates in large-scale video corpus. SR-PE is a novel pattern evaluation technique capable of measuring the spatial regularity of matching patterns formed by local keypoints. More importantly, the coherency of patterns and the perception of visual similarity, under the scenario that there could be multiple ND regions undergone arbitrary transformations, respectively, are carefully addressed through entropy measure. To demonstrate our work in large-scale dataset, a practical framework composed of three components: bag-of-words representation, local keypoint matching and SR-PE evaluation, is also proposed for the rapid detection of near-duplicates.**

***Index Terms*—Keypoints, near-duplicate detection, pattern entropy (PE), visual vocabulary.**

I. Introduction

THE choice of adopting global or local features has long been an issue of research in the field of image and video retrieval. Recently, keypoints (interest point) and the associated local features have attracted numerous attentions for their capability of characterizing salient regions which are invariant to certain geometric and photometric transformations [1], [2]. Over the past few years, keypoint-based approaches have shown success in object categorization [3], duplicate copy detection [4], semantic concept detection [5], and video search [6]. Fig. 1(a) shows an example of keypoints detected in an image pair. The keypoints jointly describe the image content by characterizing the parts which are salient and informative enough to tolerate possible transformations.

There are two general approaches to assess image content with keypoint descriptors: 1) bag-of-words (BoW) representation [6] and 2) keypoint matching [1]. BoW views keypoints as a collection of words that depict the appearance of images. To enable this representation, a codebook is generated by the offline quantization of local descriptors. Subsequently, keypoints are indexed according to the codewords found in the codebook. Bin-to-bin comparison of codewords can then be conducted to measure the similarity of images. In contrast to BoW, keypoint matching treats each image as a set of points and adopts direct point set mapping to measure similarity. In other words, the similarity of two images is based upon the degree of match between two sets of keypoints. Typically, the exhaustive search of points is required to locate the best possible matches. Because the number of keypoints can range from hundreds to thousands in an image, direct keypoint matching is much slower than BoW. Nevertheless, as no quantization loss is involved, the outcomes are more reliable than BoW.



(a) (b)

Fig. 1. ND evaluation by keypoint detection and matching. (a) Detection; (b) matching.

Keypoint matching has been demonstrated to be particularly useful for near-duplicate retrieval and detection despite of the computational issue [4], [7]–[9]. Duplicate detection can be regarded as a problem of binary decision. Specifically, given the matches found in two images (see Fig. 1), an answer of "yes" or "no" is given to confirm the duplicate identity. A typical criterion for the gating decision is to threshold the number of keypoint matches [4] or to verify by estimating the underlying geometric transformation [7]. In this paper, we propose a novel approach to determine the near-duplicate identity by assessing the matching patterns formed by two sets of keypoints. We argue that the patterns of near-duplicates are coherently regular and smooth under some unknown transformations, compared to the patterns formed by random matching. The major contributions of our work include the following.

- *Matching pattern*. We explore the coherency of patterns formed by keypoint matching of near-duplicates. A novel measure, namely Scale-Rotation invariant Pattern Entropy (SR-PE), is proposed for capturing the unique matching

Manuscript received September 20, 2007; revised October 09, 2008. Current version published January 09, 2009. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118905). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

The authors are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: cwngo@cs.cityu.edu.hk).

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/TIP.2008.2008900

patterns. The measure takes into account the transformations of near-duplicate regions in images, resulting in a robust detection algorithm even when multiple regions are separately duplicated. SR-PE is based on our initial work Pattern Entropy (PE) in [10]. Compared to PE, SR-PE shows several desirable properties for near-duplicate detection in both theoretical and practical aspects which will be further elaborated in the later sections.

- *Rapid filtering*. Keypoint matching is inherently slow particularly for large dataset. We propose a two-stage detection scheme by utilizing BoW for fast filtering in the first stage. A small pool of candidates is retained for further investigation by keypoint matching and SR-PE in the second stage. The scheme significantly speeds up the detection for several hundred times, yet without apparent degradation in performance. We apply the technique for the fast discovery of near-duplicate keyframes in one month's broadcast videos from CNN and ABC, and demonstrate very encouraging experimental results.

The definition of visual near-duplicate is generally subjective. To a certain extent, the definition does depend on the type of application being concerned. In this paper, we treat near-duplicates as the alternations of images from the same sources of image or scene. Typical example includes two images captured from the same scene but with different cameras and, thus, with the variations of viewpoint, scale, color, and lighting (see Fig. 7 for reference). Editing effects such as the insertion of logo and caption may be further added to alter their appearance. These types of near-duplicate are frequently found in the broadcast video domain, which are used to verify the performance of our proposed approaches.

The remaining sections are organized as follows. Section II briefly reviews the current literature in keypoint-based near-duplicate detection. Section III presents PE, while Section IV proposes SR-PE for evaluating the keypoint matching patterns. Section V presents our framework for the fast detection of near-duplicates in broadcast domain. Section VI shows the experimental results, and, finally, Section VII concludes our findings in this paper.

II. Related Work

Duplicate or near-duplicate detection has recently emerged as a timely research issue due to the speedy ways of capturing, duplicating and delivering videos in Internet. One important application is the detection of unauthorized use of videos and images for copyright enforcement [4], [11]. Near-duplicate detection has also become important in content-based retrieval for its potential in large-scale and web-scale search, by being able to thread near-duplicates for video ranking and summarization [12]. Depending on application, generally there could be two major categories of approaches in near-duplicate identification. One type demands speedy response while the other emphasizes detection effectiveness. A typical example of the first categories includes fingerprint-based approaches which adopt low-level global features for rapid retrieval [13]. Such approaches are only suitable for identifying duplicates in simple formatting modification such as coding and image resolution changes. In contrast, the second category of approaches often involves the intensive use of visual feature matching techniques at the image region level [4], [14] using local features such as based on keypoints. These approaches are more robust, though the robustness comes with the expense of computational cost.

In this section, we mainly review the recent works of keypoint-based approaches, which fall in the second category of near-duplicate identification. We begin by briefly introducing the popular detectors and descriptors for local keypoints, and then describe the state-of-the-art techniques in keypoint-based near-duplicate detection.

A. Keypoint Detector and Descriptor

There are numerous keypoint detectors and local descriptors in the literature [2], [15]. The detectors basically locate stable keypoints (and their support patches) which are invariant to certain variations introduced by geometric and photometric changes. Popular detectors include Harris-Affine [16], Hessian-Affine [16], Difference of Gaussian (DoG) [1], and Maximal Stable Extreme Region (MSER) [17].

To match keypoints across two images (as illustrated in Fig. 1), one factor is to have robust features for describing local patches surrounding keypoints. Generally the descriptors can turn up quite different performance in response to various transformations. A recent study in [15] shows that SIFT [1] family can always attain satisfactory performance under different contexts. SIFT is typically a 3-D histogram of gradient location and orientation, by quantizing the local patch of keypoint to 4×4 grids. To date, there are several variants of SIFT. Among them, PCA-SIFT and GLOH (Gradient Location-Orientation Histogram) [15] are the most representative ones. PCA-SIFT [18] performs PCA (Principal Component Analysis) on the gradient field of local patch. GLOH is similar to SIFT except the quantization of local patch is based on log-polar grid.

B. Keypoint-Based Near-Duplicate Detection

Existing detection approaches based on keypoints include [4], [7], [8], [10], and [14]. In [14], keypoints in an image are modeled as a stochastic attributed relational graph (ARG) and near-duplicates are detected by graph learning and matching techniques. In [4], locality sensitive hashing (LSH) is adopted for fast searching of similar keypoints. In [8], a fault-tolerant matching scheme called one-to-one symmetric (OOS) is proposed for reliable keypoint matching. With this scheme, a filtering index LIP-IS is further adopted for rapid searching of the nearest keypoints. Empirically, the OOS plus LIP-IS scheme shows better performance than LSH. Similarly, in [7], a distortion-based probability similarity search algorithm is proposed for fast duplicate retrieval with keypoints. This piece of work is further extended in [19] to allow keypoint trajectories being tracked and indexed to support the localization of duplicate video segments.

There exist several strategies to determine the near-duplicate identity, once after the keypoints between two images are matched. Existing strategies can be broadly categorized as cardinality threshold (CT) and pattern learning (PL) based evaluation. A common strategy for CT-based evaluation is to

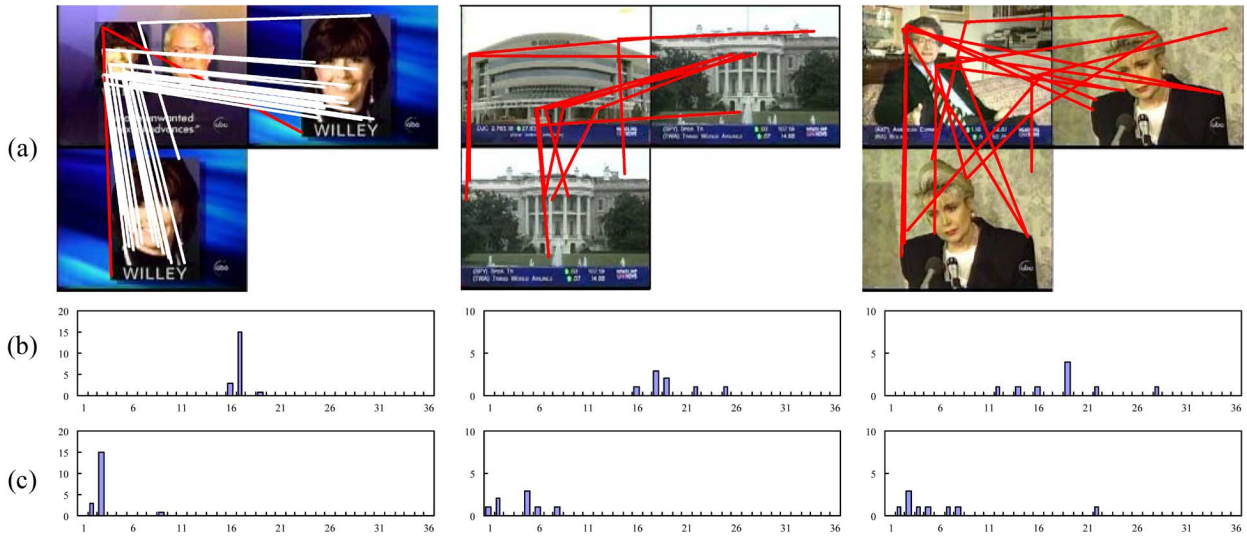


Fig. 2. Histograms of matching patterns for three ND pairs: (a) matching lines between keypoints, (b) vertical histogram $\mathcal{G}_v$, and (c) horizontal histogram $\mathcal{G}_h$ (NOTE: In this and the remaining figures, red lines indicate the false keypoint matches, while lines in other colors mean correct matches).

simply threshold the cardinality of keypoints being matched [1], [4]. The threshold setting, nevertheless, is sensitive owing to the fact that the keypoint cardinality is often dependent on scene complexity. Typically the number of keypoints being matched in an image can range from a few to several thousands, resulting in unstable performance. A more robust CT-based approach is to perform geometric consistency checking by employing RANSAC (RANdom SAmple Consensus) or Hough transform [1], [4], [7], [19]. Thresholding the number of matched keypoints obeying geometric consistency is still necessary to determine the near-duplicate identity. In addition, RANSAC is only applicable when the correct matches dominate false matches. Hough transform, on the other hand, favors images with certain regular shapes like lines and circles. In general, these strategies face limitation when duplicate objects are embedded in highly cluttered and complex background. More sophisticated approaches are to take into account the neighborhood configuration for mining near-duplicate region from noisy matches. In [6] and [20], spatial consistency is enforced by checking whether the matches are consistent within a predefined neighborhood area. The consistency could be based on either the number matches or the spatial layout of neighboring matches. To be efficient, these approaches adopt bag-of-words (BoW) representation for matching keypoints. BoW is known to be suffered from quantization error. The recent work in [21] and [22] thus enhances BoW by using Latent Dirichlet Allocation (LDA) to robustly reduce false matches while considering neighborhood constraint. This work [21], [22] is shown to be effective for locating the key-places, of very often exhibiting weak overlap of duplicate scene, within a movie.

PL-based evaluation emphasizes the learning of near-duplicate matching patterns. These matching patterns are assumed to follow certain spatial regularity. Without explicitly stating how the spatial regularity is determined under a predefined neighborhood configuration as in [6] and [20]–[22], PL-based strategy evaluates likelihood of near-duplicate based on the matching pattern itself. In addition, contrary to [6] and [20]–[22] which perform matching based on BoW representation for efficiency purpose, PL-based strategy allows consideration of keypoint-level matching which is more accurate than BoW. Our recent work in [8] adopts a learning based approach by training from the matching patterns of near-duplicate and random pairs with Support Vector Machines (SVM). An unsupervised learning approach based on pattern entropy (PE) is also proposed in [10]. PE, nevertheless, has several deficiency by assuming only the commonest and easiest matching patterns. In this paper, we further extend PE to SR-PE for more stable and reliable near-duplicate detection.

III. Entropy-Based Evaluation

By definition, near-duplicate (ND) pairs share the duplication of regions (either in background scene, objects or both). Having the reliable local descriptors, the upshot of keypoint matching ideally should form certain spatially coherent patterns that are different from random matching. These patterns could be one or several bunches of parallel or zoom-like matching lines across the sub-regions of keypoints. Fig. 2(a) shows the examples of matching patterns for ND (left) and non-ND (middle and right) pairs. The matchings of non-ND pairs often show random patterns with matching lines being arbitrarily crossed across space. The matching formats provide vivid pattern cues for the discrimination of ND and non-ND pairs. Based on this observation, we describe a measure, called pattern entropy (PE), to evaluate the information of being an ND pair.

PE captures the matching patterns with two histograms of matching orientations. The histograms $\mathcal{G}_h$ and $\mathcal{G}_v$ are constructed, respectively, by aligning two images horizontally and vertically, as shown in Fig. 3. Depending on the alignment, a histogram is composed of the quantized angles formed by the matching lines and horizontal or vertical axis. Denote h as the height of the upper image (image 1 in Fig. 3), and the coordinates of keypoint A in image 1 and keypoint A' in image 2 as (x_0, y_0) and (x_1, y_1), respectively. The angle θ_v of a line

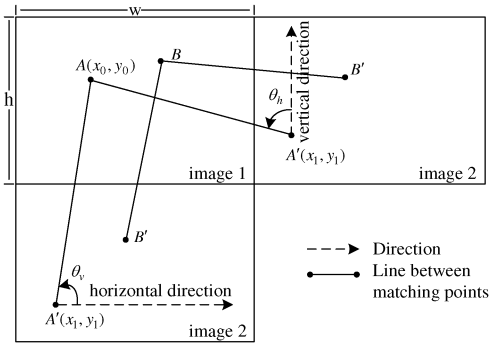


Fig. 3. Computing the orientation of matching lines by the horizontal and vertical alignment of two images.

formed by two matched keypoints is computed as

$$\theta_v = \arccos\left(\frac{x_1 - x_0}{\sqrt{(x_1 - x_0)^2 + (y_1 + h - y_0)^2}}\right). \quad (1)$$

The histogram $\mathcal{G}_v$ is formed by computing θ_v of lines and then accumulating the count to the corresponding bins. The histogram $\mathcal{G}_h$ is computed in a similar manner by the angle θ_h. Denote w as the width of left image (image 1), the angle θ_h is computed as

$$\theta_h = \arccos\left(\frac{x_1 - x_0}{\sqrt{(x_1 + w - x_0)^2 + (y_1 - y_0)^2}}\right). \quad (2)$$

The histograms is quantized into 36 bins with a step of 5° from 0° to 180°. Ideally the parallel or nearly parallel lines should fall in the same bin of a histogram. The histograms $\mathcal{G}_h$ and $\mathcal{G}_v$ intuitively hint the spatial coherency of matching in the horizontal and vertical directions. Fig. 2(b) and (c) shows the histograms corresponding to an ND (left) and two non-ND pairs (middle and right). The distributions of $\mathcal{G}_h$ and $\mathcal{G}_v$ depict different partitions of orientation for the same set of matched keypoints. For an ND pair, both histograms should be correlated. Specifically, whenever a peak in $\mathcal{G}_h$ is found, there exists a corresponding peak of the same keypoints in $\mathcal{G}_v$. To reveal the mutual information between $\mathcal{G}_h$ and $\mathcal{G}_v$, PE uses entropy to measure the homogeneity of histogram patterns.

Denote N as the number of bins in a histogram, and define two order sets $\mathbf{P} = [p_1, p_2, \ldots, p_m]$ and $\mathbf{Q} = [q_1, q_2, \ldots, q_n]$, where $m \leq n \leq N$. The sets $\mathbf{P}$ and $\mathbf{Q}$ collect keypoints that fall in the nonempty bins of $\mathcal{G}_h$ and $\mathcal{G}_v$, respectively. The notation p_i (similarly for q_i) represents a nonempty set of keypoint pairs in a bin of histogram. Physically, $\mathbf{P}$ corresponds to one of the histograms ($\mathcal{G}_h$ or $\mathcal{G}_v$) with less nonempty bins, while $\mathbf{Q}$ corresponds to the other histogram. In principle, $\mathbf{P}$ is more compact than $\mathbf{Q}$ since less bins are used to accommodate the matched keypoints. PE measures the degree of points in p_i being distributed in $\mathbf{Q}$, defined as

$$\mathrm{PE}(\mathbf{Q}, \mathbf{P}) = \frac{1}{M} \sum_{q_i \in \mathbf{Q}} \mathrm{Entropy}(q_i, \mathbf{P}) \quad (3)$$

where

$$\mathrm{Entropy}(q_i, \mathbf{P}) = -\frac{1}{\log m} \sum_{p_j \in \mathbf{P}} \frac{|q_i \cap p_j|}{|q_i|} \times \log \frac{|q_i \cap p_j|}{|q_i|} \quad (4)$$

$$M = \sum_{q_i \in \mathbf{Q}} |q_i| = \sum_{p_i \in \mathbf{P}} |p_i| \quad (5)$$

$|p_i \cap q_i|$ is the cardinality of intersection between two sets p_i and q_i, and M is the total number of matching lines. Basically, $\mathrm{Entropy} = [0, 1]$ measures the extent of dispersing a set p_i across the bins of another histogram of the orthogonal direction. An entropy value of 0 indicates the keypoints in p_i are found exactly in another set q_i of $\mathbf{Q}$. A value of 1 indicates that the keypoints in p_i are evenly distributed in some sets of $\mathbf{Q}$. The range of $\mathrm{PE} = [0, 1]$. The extreme value of $\mathrm{PE} = 0$ indicates a perfect coherent match in both horizontal and vertical direction. Conversely, $\mathrm{PE} = 1$ basically hints a random match across space. Note that in (3), practically p_i with low cardinality should be excluded from pattern evaluation. First, bins containing few keypoints may be caused by problems such as quantization error. Second, these bins can form arbitrary matches which distract the accuracy of PE measure. For this purpose, a parameter γ is required to gate whether a set p_i should participate in PE evaluation [i.e., (3) is computed based upon all $p_i \geq \gamma$]. In latter experiments, we will further discuss the sensitivity of the parameter γ.

The PE bears two interesting facts. The ND and non-ND pairs can be distinguished according to the degree of matching coherency (and randomness) in space formed by keypoints. Secondly, the measure fits the symmetric property of ND pairs due to the selection of $\mathbf{P}$ for testing the dispersion of p_i with respect to q_i. PE can be considered as a novel version of cross-bin histogram matching with many-to-many mapping such as EMD [23], where the "weight" of each matching unit is characterized by keypoints while the flow is the number of keypoints transmitted from one unit to the other. However, unlike EMD which measures the distance of every match, PE considers the dispersion of matches across bins.

IV. SR-Pattern Entropy

Given an ND pair, there could exist one or more than one or multiple near-duplicate regions probably undergone geometric transformations such as scaling and rotation. Specifically, each matched region pair is transformable from one to another under arbitrary combination of geometric operations. PE, utilizing the histogram of matching orientations, is excellent for characterizing regions with parallel-like matching lines, but not for regions undergone considerably scale and rotation changes. For scaling effect, the matching lines of a region pair could be in different orientations and, thus, are distributed in different bins of histogram. As a consequence, these matches might probably be regarded as noises and not considered during PE evaluation. For rotation effect, it is possible that the matching lines of a

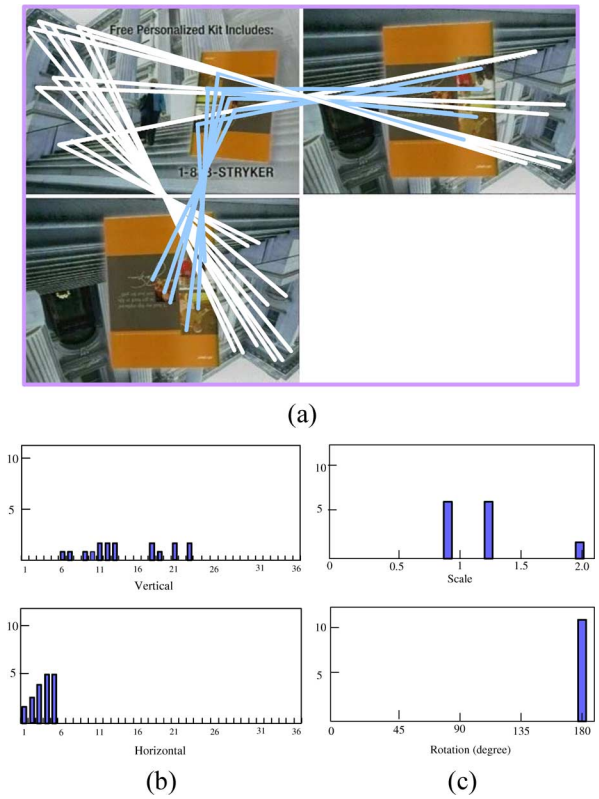


Fig. 4. Representing the matching lines of an ND image pair (a) by using (b) PE: vertical and horizontal histograms of matching orientations; (c) SR-PE: clusters of ND regions in scale and rotation channels.

region pair cross each other, resulting in difficulty of distinguishing them from random matching patterns. Fig. 4(a) depicts the matching lines of an ND image pair which undergoes scaling and rotational changes. Fig. 4(b) shows the corresponding vertical and horizontal histograms. As noted, there is no apparent peak observed in the vertical histogram. When performing PE evaluation, the bins in two histograms are not coherently matched, resulting in high entropy value.

In this section, we propose a revised version of PE, namely Scale and Rotation invariance PE (SR-PE), for more robust pattern evaluation. In addition to the consideration of scaling and rotational effects, SR-PE adopts mean-shift [24] for clustering of matching lines, rather than histograms as used in PE, to minimize the impact of quantization error. Fig. 5(c) illustrates the outcome of SR-PE, which produces more compact representation of matching lines than PE in Fig. 5(b), when considering the transformation of ND regions. More details will be described in Section IV-B.

A. *Identification of Near-Duplicate Regions*

Denote $\mathcal{R}_i$ and $\mathcal{R}_j$ as a pair of near-duplicate regions, the transformation between them can be expressed as

$$\mathrm{R}_j = s \times \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \times \mathrm{R}_i \qquad (6)$$

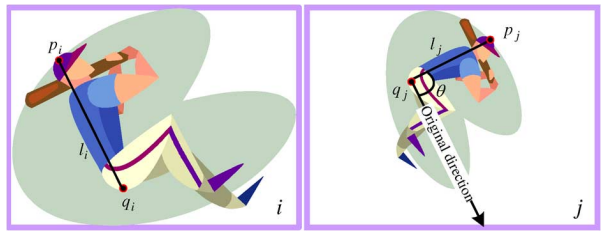


Fig. 5. SR-PE: estimating rotation and scale changes between image i and image j from two pairs of matched points (p_i, p_j) and (q_i, q_j).

where the parameters s and θ represent the scale and rotation, respectively. SR-PE aims to cluster the matching lines according to their underlying region pairs, under the scenario that the ND regions and their transformed parameters are unknown. To solve this problem, we estimate the scale s and rotation θ of every two matching pairs as observation. The ND regions are then approximated by treating the observations as features for clustering.

Take Fig. 5 as reference, the regions are arbitrarily scaled, rotated, and translated. Let p_i and q_i as two keypoints of a region in image i, and are matched to p_j and q_j of image j, respectively. We denote the length l_i as the Euclidean distance from p_i to q_i, and similarly for the length l_j. The scale s of two matching lines are estimated with l_i/l_j, the rotation θ is the angle between l_i and l_j when superimposing both images. In brief, by computing the s and θ of every two matching pairs, the group of matching lines undergone similar transformation in a ND region pair ideally can be discovered through clustering.

Algorithm (1) summarizes the procedure of identifying ND region pairs in two images. In *Step*-1, an exhaustive estimation of transformation parameters for all pairs of matching lines is carried out. The parameters are clustered with mean-shift algorithm [24], respectively, in the scale s and rotation θ channels. Mean-shift algorithm is popular in both computer vision and data mining communities for its versatility and simplicity. In the algorithm, each cluster in a channel corresponds to one ND region pair. Because each matching pair formed by (p_i, p_j) involves multiple estimations of s and θ in *Step*-1, each (p_i, p_j) basically can be clustered into more than one group. *Step*-3 defuzzes the cluster membership of (p_i, p_j) by assigning it to the most likely cluster where (p_i, p_j) resides most of the times. Note that Algorithm (1) produces two sets of region pairs corresponding to scale and rotation channels, respectively.

Algorithm 1: Estimating near-duplicate region pairs

Input: The set of matched keypoints between 2 images (i, j)

Output: Near-duplicate region pairs

Step 1:

for each pair of matching keypoints (p_i, p_j) **do**

for each remaining pair (q_i, q_j) **do**

Estimate scale s and rotation θ based on (6)

end for

end for

Step 2: Mean shift clustering in s and θ channels
Step 3: Defuzzing the membership of each (p_i, p_j) pair in both s and θ channels

In Algorithm (1), each matched pair (p_i, p_j) takes multiple reference points for estimating the transformation parameter. Let n as the number of matching lines, *Step*-1 involves $n \times (n-1)/2$ estimations. Based on mean-shift clustering, the complexity of Algorithm (1) is thus $O(n^2 \log(n))$. The complexity can indeed be reduced by an order of magnitude to $O(n \log(n))$, by simply taking only few spatially closed matching keypoints as references in *Step*-1. The associated risk is that the estimation could be noisy if errors happen in the keypoint detection and matching steps. To compromise, we consider three reference points q_i, referring to the three most spatially nearest neighbors, in *Step*-1. In principle, at least two points are required for computing (6), whether using more reference points is a tradeoff between speed and accuracy. In our case, to safeguard the performance, this heuristic strategy is adopted only when $n > 40$. When the value of n is large, the chance of ND is relatively higher, and, thus, using brute-force estimation in *Step*-1 is, indeed, not necessary.

B. Near-Duplicate Detection

The ND region identification in Algorithm (1) produces two partitions of region pairs, respectively, in the scale and rotation channels. Similar to the histogram bins of pattern entropy (PE) in (3), each partition reveals how the region pairs match and distribute in a channel according to their underlying transformations. Moreover, one cluster in a partition is like a nonempty bin of histogram in one direction. Similar to the notation in (3), we denote the two partitions as $\mathbf{P} = \{p_1, p_2, \ldots, p_m\}$ and $\mathbf{Q} = \{q_1, q_2, \ldots, q_n\}$, where p_i and q_i represent clusters and $m \leq n$. The partition $\mathbf{P}$ is either from the scale or rotation channel whichever with a smaller number of cluster numbers. Given the two partitions $\mathbf{P}$ and $\mathbf{Q}$ in scale and rotation channels, (3) is employed directly to compute the pattern entropy PE_{sr} in the same fashion. Similar to PE, practically a parameter γ is also set to gate whether a cluster should involve in PE_{sr} evaluation, depending on the size of the cluster. Generally, clusters with smaller size could probably due to noise and, thus, should be discarded from evaluation.

Fig. 4 gives an example to illustrate the robustness of SR-PE compared to PE. Fig. 4(c) shows the distributions of clusters in scale and rotation channels, respectively. Compared to the histograms by PE in Fig. 4(b), there are obvious peaks in both scale and rotation channels corresponding to the underlying transformation of ND regions of the image pair shown in Fig. 4(a). These peaks are correlated to each other across channels and are matched to produce a lower entropy value of PE_{sr}, compared to PE which could not distinguish the pattern under transformation from random matching.

Although PE_{sr} is capable of capturing the homogeneity of ND region pairs across scale and rotation, it cannot distinguish the case when only few matching lines are found in each region pair. Fig. 6 illustrates the problem case where the toy examples (left column) in Fig. 6(a)–(c) have the same entropy value $(\mathrm{PE}_{sr} = 0)$, although the likelihood of being near-duplicate is much higher for Fig. 6(c) compared to Fig. 6(a) and (b). To take this into account, SR-PE considers not only the homogeneity but also compactness of partitions. Let a partition as $\mathbf{G} = \{g_1, g_2, \ldots, g_m\}$ with m clusters, we employ entropy to measure the compactness. The compactness of the scale channel is computed as

$$\mathrm{PE}_s = -\frac{1}{logm} \sum_{g_i \in G} \frac{|g_i|}{z} \times \log \frac{|g_i|}{z} \tag{7}$$

where

$$z = \sum_{g_i} |g_i|. \tag{8}$$

The compactness of rotation channel PE_r is computed in the similar way. Thus, in addition to PE_{sr}, SR-PE includes two other measures PE_s and PE_r for entropy-based evaluation. To combine these three measures, we adopt a simple nonlinear fusion as follows:

$$\mathrm{SR-PE} = \max\{\mathrm{PE}_{sr}, \mathrm{PE}_s, \mathrm{PE}_r\} \tag{9}$$

which picks the measure with the highest entropy value. In other words, SR-PE gets a lower entropy value indicating perfect match only when the ND region pairs show homogeneity across channels and compactness within channels.

Comparing with PE, SR-PE offers several significance. First, the matching pattern is determined based upon the assumption that the underlying transformation is unknown. Thus, the matching lines of a region pair even crossing each other or in star-like pattern can still be recognized by SR-PE as long as they follow certain transformation. Second, SR-PE enlightens the possibility of locating near-duplicate regions while PE basing on histogram cannot deal with localization effectively. Third, SR-PE considers the perception of visual near-duplicate. More specifically, the incremental changes of near-duplicate judgment is taken into account, as demonstrated in Fig. 6, when more and more matching lines are found in the ND region pairs.

V. Application to Fast Detection of Near-Duplicate Keyframes

An emerging application of ND detection is to automatically thread near-duplicate keyframes in large video corpus to facilitate search and browsing [12]. In broadcast video domain, there are abundant near-duplicate keyframes found across different channels and time zones [10]. Chaining these ND keyframes can greatly facilitate the clustering and organization of news topics evolving over times. Nevertheless, the automatic discovery of ND keyframes in a large corpus remains an impractical problem even with the good detection accuracy offered by matching local descriptors [10], [14]. Considering a video corpus containing 7,000 keyframes which can generate more than twenty million pairs of keyframes, locating ND pairs can be extremely time consuming. Even with the index structure such as locality sensitive hashing [4], the matching of keypoints for all candidate pairs can still take several days. Basically, given n keyframes, the amount of candidate pairs grows in $O(n^2)$, which makes the

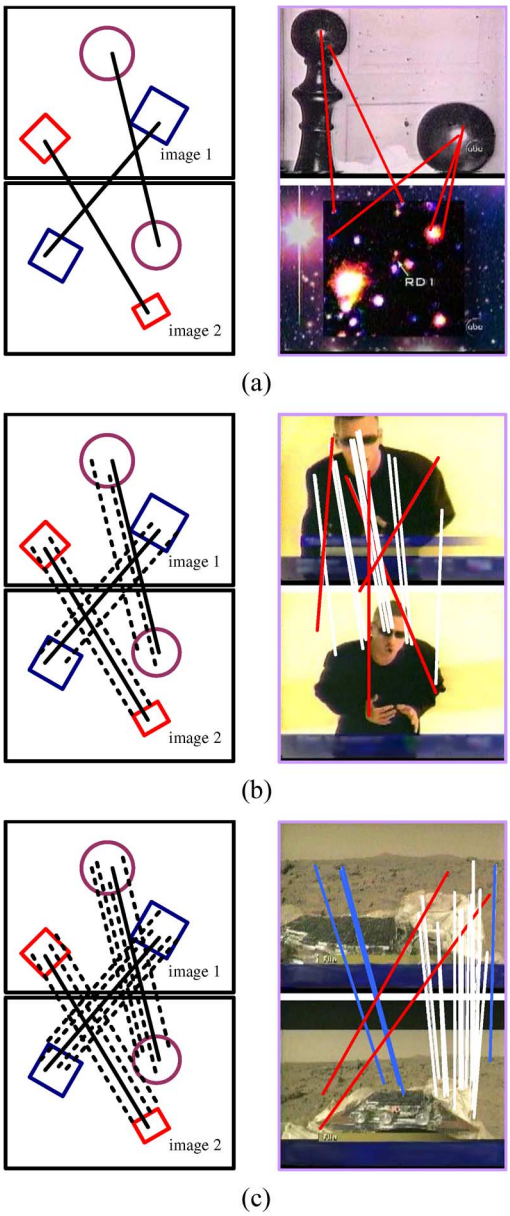


Fig. 6. Incremental change of near-duplicate judgment when more matching pairs are found: (left) toy examples, (right) real image pairs. The examples in (a)–(c) have similar PE_{sr} value, but different visual impression of ND when more matching lines are found. Note that (a) is not an ND pair. (a) Approximately one matching line per ND region; (b) few more matching lines; (c) more matching lines.

exhaustive checking of all possible candidate pairs an infeasible solution.

Given n keyframes, an apparent observation is that the growing of non-ND pairs is much more than ND pairs. By a larger value of n, typically non-ND candidates increase in quadratic speed, while ND candidates grow in linear or sub-linear speed. This actually hints the necessity of adopting a filtering scheme to retain as few as candidate pairs for keypoint matching and pattern evaluation. Basically given $O(n^2)$ candidate pool, we only need to perform $O(kn)$ checking, supposing only k candidates are retained for each evaluation. For this reason, we propose a filtering scheme by offline quantizing the keypoint descriptors into a codebook or visual vocabulary. Each keyframe is then coded as a bag-of-words with the vocabulary. Instead of performing keypoint matching, the initial stage is to retrieve a small pool of candidates of every keyframe for detailed evaluation. Exhaustive matching and pattern evaluation are then conducted to investigate the degree of near-duplicate of each candidate pair.

The bag-of-words representation has been applied in [6]. The retrieval speed can be extremely fast due to the nature of sparse vector representation and the use of inverted file index [6]. In our case, we employ k-means algorithm to quantize the keypoint descriptors into thousands of clusters. The centroid of each cluster corresponds to a visual keyword. When assigning words to a keyframe, the corresponding descriptors are directly matched to the nearest words. The importance of a word is then weighted with term frequency (tf), which describes the number of words appeared in a keyframe. With this vector representation, retrieving ND keyframes on a database of more than one thousand hours of videos can be completed as fast as within a second. By retaining only the top-k keyframes for further investigation, the detection of ND pairs is expected to be extremely efficient.

VI. Experiments

In this section, we begin by introducing the datasets (Section VI-A) and keypoint descriptors (Section VI-B) used in the experiments. Then two main experiments are presented to justify the performance of SR-PE with comparison to other existing approaches. These experiments examine several aspects of ND detection, including performance comparison of SR-PE with other approaches (Section VI-C), and efficiency of using VK filtering prior to SR-PE (Section VI-D) for fast ND detection. All the experiments are conducted on a 3-GHz Pentium-4 machine with 512-M memory. The algorithms are implemented with C++ and compiled by GCC 3.4.2 (mingw-special).

A. Dataset and Performance Evaluation

We use two datasets from: 1) Columbia [25], 2) CityU [26] of different sizes for experiments. Both datasets are separate subsets selected from TRECVID 2003 video corpus [27]. The first dataset consists of 600 keyframes with 179,700 candidate pairs. Among them, 210 pairs are near-duplicate. The second dataset spans one month's broadcast videos and, thus, is more challenging. It covers 52 broadcasts of CNN and ABC channels in March of 1998, with approximately two news reports per day. These videos have a total of 29 news topics covering 805 stories and resulting in 7,006 shots. One representative keyframe per shot, as specified in TRECVID, is selected for experiments. As a consequence, these keyframes form 24,538,515 candidate pairs for detection. Among them, only 3,388 pairs are near-duplicate copies. These pairs form 693 ND threads and involve a total of 1,953 keyframes. Considering the amount of ND pairs against the candidate pool, the chance of finding a correct pair in random is as low as $1.38e^{-4}$.

In CityU's dataset, three assessors were involved in labeling the ground-truth. Table I lists the top five most frequently reported news themes, along with their amount of ND pairs, in the ground-truth. Fig. 7 also lists few examples of challenging ND

TABLE I
TOP FIVE FREQUENTLY REPORTED NEWS TOPICS IN CNN AND ABC CHANNELS DURING THE WHOLE MONTH OF MARCH 1998

Theme	# of Stories	# of ND Pairs
Basketball	78	73
Clinton Sexual Scandal	58	250
Finance	53	159
El nino	38	12
Arkansas school shooting	37	149

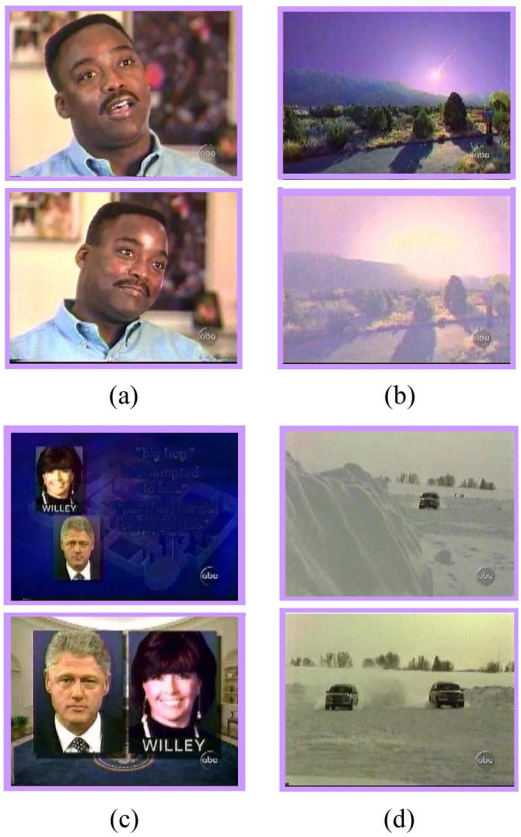


Fig. 7. Examples of near-duplicates in CityU dataset: (a) different acquisition time, (b) illumination change, (c) multiple duplicate regions, and (d) viewpoint and scale variation.

pairs in this dataset. To avoid the potential ambiguity in identifying ND pairs, two assessors were asked to label the dataset separately. The third assessor compared their results, compiled and then grouped them into one final ground-truth set. Careful, although subjective, judgment was required for the third assessor whenever there was a conflict of an ND pair. In Columbia's dataset, there are originally 150 ND pairs. However, after the careful evaluation by three assessors, another 60 pairs are located from the remaining 300 non-ND keyframes. The new groundtruth with 210 ND pairs can be found at [26].

To evaluate the performance of ND detection, we use Recall, Precision and F-measure, defined as

$$\text{Recall} = \frac{\text{Number of ND pairs correctly detected}}{\text{Total Number of ND pairs}} \quad (10)$$

$$\text{Precision} = \frac{\text{Number of ND pairs correctly detected}}{\text{Number of detected ND pairs}} \quad (11)$$

$$\text{F}-\text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (12)$$

Recall measures the accuracy of returning ground-truth ND pairs, while Precision assesses the ability of excluding false positives. F-measure calculates the fitness of ground-truth and detected ND pairs by jointly considering recall and precision.

B. Feature Description

We employ Hessian-Affine [28] for keypoint detection in the experiments. In our previous studies [10], [29], Hessian-Affine, compared to the more popular detector DoG (Difference of Gaussian), on average locates less number of keypoints per keyframe, and yet shows similar retrieval performance as DoG. To describe keypoints, we use a newly proposed descriptor, namely PSIFT, for experiments. PSIFT is a PCA version of SIFT [1] in lower dimension. Note that this descriptor is different from PCA-SIFT [18] which performs PCA on the gradient field of a local patch. The eigenspace of PSIFT is offline precomputed based on SIFT descriptors extracted from a training set. Empirically we retain the first 36 eigenvectors, which correspond to the first 36 eigenvalues sorted in descending order, of SIFT descriptors for linear projection. The 36 eigenvectors basically capture the major variances of SIFT. As a consequence, PSIFT, as PCA-SIFT, is also a descriptor in 36-dimensional space.

In addition to using PSIFT as the descriptor for ND detection, PSIFT is also employed to generate a codebook of visual keywords for ND filtering. Due to the limitation of space, we do not show the detailed performance of PSIFT. Basically, PSIFT has similar performance as SIFT, and better performance than PCA-SIFT and GLOH. Comparing to SIFT which is originally in 128 dimensions, PSIFT with feature length of 36 dimensions offers several advantages. During ND filtering, the assignment of keypoints to visual keywords, which involves feature comparison, becomes efficient due to the reduction in feature dimension. More importantly, ND detection, which is also a process of feature comparison, could be more efficiently conducted in lower dimensional space, particularly with the use of indexing structure such as LIP-IS in our case.

C. ND Detection: Performance Comparison

In this experiment, the ND detection is based on the brute-force (exhaustive) search of near-duplicate pairs. We compare six approaches: SR-pattern entropy (SR-PE), pattern entropy (PE), cardinality threshold (CT), RANSAC-based cardinality threshold (RS-CT), visual keyword (VK) and block-based color moment (CM). SR-PE and PE adopt LIP-IS for fast keypoint filtering and OOS for matching [10]. CT and RS-CT are based on the work in [4] by adopting nearest neighbor search to match keypoints. In CT and RS-CT, a gating threshold (η), which is the cardinality of matched keypoints, is set for determining ND pairs. In principle, a candidate should have a large enough number of keypoints being matched in order to declare the ND identity. RS-CT is basically a robust version of CT by employing RANSAC to prune noisy matches before thresholding. VK and CM serve as baselines to judge the improvement of three other approaches. Both features, basing on the histogram of visual keywords and colors, respectively, rank the ND candidate pairs in descending order according to their similarity.

TABLE II
ND DETECTION: PE—PATTERN ENTROPY, CT—CARDINALITY THRESHOLD, VK—VISUAL KEYWORD, CM—COLOR MOMENT (THE BEST RESULTS ARE BOLD)

	Columbia Dataset			CityU Dataset		
	Prec	Rec	F-M	Prec	Rec	F-M
SR-PE	1.0	0.824	**0.903**	0.91	0.784	**0.842**
PE	0.977	0.810	0.885	0.907	0.769	0.832
RS-CT	0.912	0.791	0.847	0.873	0.748	0.806
CT	0.961	0.709	0.816	0.868	0.740	0.799
VK	0.593	0.593	0.593	0.372	0.372	0.372
CM	0.4	0.4	0.4	0.192	0.192	0.192

(Prec: Precision, Rec: Recall, F-M: F-Measure)

Table II shows the performance comparison of six different approaches in Columbia and CityU datasets. For SR-PE, PE, CT and RS-CT, the corresponding parameters (γ for SR-PE and PE, and η for both CT and RS-CT) are empirically set to obtain the best possible performance. For VK and CM, the top-k similar pairs are declared as ND pairs, where the value of k is set equal to the number of groundtruth ND pairs in the datasets. As indicated in the tables, keypoint based approaches (SR-PE, PE, CT, RS-CT, VK) outperform color moment (CM) significantly, particularly when the candidate pool increases and reaches to more than twenty million as in CityU dataset. In addition, the approaches with matching strategy (SR-PE, PE, CT, RS-CT) are also constantly better than codebook based comparison (VK). The results are not surprised since keypoint matching strategy does not involve quantization loss as in VK, while the feature is more tolerant to geometric and photometric changes compared to CM.

1) Comparing PE-Based Versus CT-Based Detection: There are two key factors determining the performances of matching-based approaches: number of matching lines, and transformations of ND regions. An interesting fact in the experiment is that by simply thresholding the number of matching lines (η), CT has already offered satisfactory performance. Nevertheless, because the number of matching lines can range from as few as five to as large as several hundreds, proper setting of η is always difficult. RS-CT successfully improves CT by also considering the dominant transformation of ND regions. However, due to the consideration of single dominant transformation, this approach is potentially limited and risky if multiple ND regions undergo different transformations.

SR-PE, in contrast to CT and RS-CT, considers the transformations of multiple regions separately and interprets the transformations as a pattern that can be rigorously measured by entropy. In addition, the compactness of regions measured via the number of matching lines is also taken into account for evaluating pattern entropy. These two characteristics actually make SR-PE a relatively robust approach compared to CT and RS-CT. Fig. 8 shows examples which are successfully detected by SR-PE while not by other approaches. Fig. 8(a) shows a non-ND pair with excessive number of matching lines, while Fig. 8(b) shows a ND pair with fewer number of matching lines. Thresholding approach such as CT performs poorly for these types of images. For RS-CT, the matching lines in Fig. 8(a) are wrongly estimated and confirmed as a transformation by RANSAC. While in Fig. 8(b), due to lack of matching lines to confirm the transformation estimated by RANSAC, the ND pair is missed by RS-CT. Fig. 8(c) further shows another interesting example where there are two ND regions, each with few matching lines (marked in white and light blue color). The images in Fig. 8(c) undergone scaling changes. But due to depth variation, the background scene exhibits different transformation from the foreground moving objects. As a consequence, by considering both regions separately, SR-PE successfully detects this example but not RS-CT.

(a) (b) (c)

Fig. 8. Examples of ND and non-ND pairs robustly detected by SR-PE but not by CT and RS-CT: (a) non-ND pair but with excessive number of matching lines, (b) true ND pair but with few matching lines, and (c) true ND pair with multiple duplicate regions.

TABLE III
PERFORMANCE OF SR-PE AND PE WITH VARYING PARAMETER SETTINGS. (A) COLUMBIA DATASET (B) CITYU DATASET

(a)

γ	SR-PE			PE		
	Prec	Recall	F-measure	Prec	Recall	F-measure
3	0.919	0.867	0.892	0.815	0.857	0.835
4	0.983	0.833	0.902	0.977	0.810	**0.885**
5	1.0	0.824	**0.903**	1.0	0.748	0.856
6	1.0	0.810	0.895	1.0	0.695	0.820
7	1.0	0.786	0.88	1.0	0.686	0.814

(b)

γ	SR-PE			PE		
	Prec	Recall	F-measure	Prec	Recall	F-measure
3	0.782	0.830	0.806	0.698	0.824	0.756
4	0.852	0.816	0.834	0.852	0.793	0.821
5	0.884	0.798	0.837	0.907	0.769	**0.832**
6	0.910	0.784	**0.842**	0.928	0.743	0.825
7	0.928	0.765	0.838	0.948	0.718	0.817

TABLE IV
AVERAGE TIME COST FOR EVALUATION OF ONE IMAGE PAIR

Method	SR-PE	PE	CT	RS-CT	VK	CM
Time (ms)	90	90	75	76	0.015	0.00009

Fig. 9 shows examples of ND pairs falsely detected and missed by SR-PE. Pairs in Fig. 9(a) and (b) are falsely detected mainly due to the somewhat similar background scenes. For example, the images in Fig. 9(b) share similar background outline in the far background. The miss detections, as shown in Fig. 9(c)–(e), are mainly due to large scene variation as a result of a combination of different effects such as viewpoint change, scaling and multiple moving objects.

2) Comparing SR-PE Versus PE: In SR-PE and PE, the only required parameter is γ. The parameter is for practical consideration in order to reduce the noise caused during clustering or his-

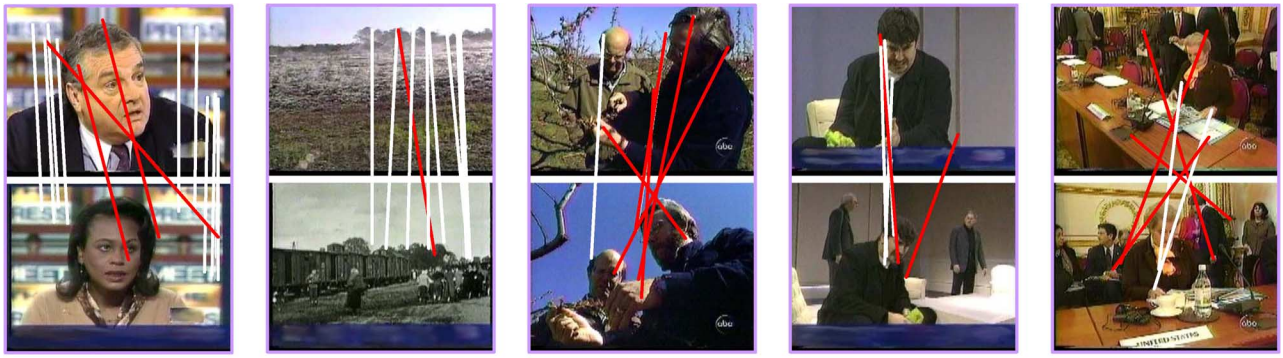


Fig. 9. Failure ND examples by SR-PE: (a)–(b) false positives and (c)–(e) false negatives.

TABLE V
VK FILTERING ON CITYU DATASET WITH DIFFERENT TOP-k

Top-k	5	10	15	20	25	30	35	40	45
Precision	0.934	**0.944**	0.943	0.940	0.937	0.935	0.932	0.931	0.929
Recall	0.402	0.554	0.635	0.679	0.705	0.721	0.729	0.736	**0.74**
F-measure	0.564	0.698	0.759	0.788	0.804	0.814	0.818	0.822	**0.824**
Number of pairs	31,598	61,294	90,026	118,188	145,975	173,497	200,696	227,651	258,552
Speed-up ratio	776	399	271	206	166	140	121	106	95

togramming of matching lines. Table III experiments the sensitivity of γ for both SR-PE and PE. In terms of F-measure, SR-PE and PE are indeed less sensitive to the setting of γ. SR-PE, in particular, maintains a relatively stable performance in terms of precision, recall and F-measure across all the settings. Basically, PE can be viewed as a special case of SR-PE. PE mainly intends to capture the parallel and zoom like patterns of the matching lines across two images. Basically, it works well when images are under slight scale and rotation transformations on the region level. SR-PE, on the other hand, is scale and rotation invariant on the region level and, thus, demonstrates constantly better performance than PE across various settings of γ.

3) Speed Efficiency: Table IV lists the average computational time of six different approaches in evaluating one pair of images. SR-PE is computationally efficient and can process more than 10 image pairs per second. Nonmatching based approaches such as VK and CM, nevertheless, are even faster by several thousand times than SR-PE. Practically, both VK and SR-PE can be integrated to leverage the accuracy and speed of matching and nonmatching based approaches.

D. Fast Detection with VK Filtering

The brute-force type of detection based upon keypoint matching is expected to be inefficient. In this experiment, we adopt a codebook of visual keywords (VK) for ND filtering and investigate its impact on CityU dataset. The codebook is generated by randomly selecting 800 keyframes from the dataset for keypoint clustering. Table V shows the performance of ND detection by using the proposed filtering scheme. We experiment the impact by varying the number of top-k candidates to be kept for further evaluation by SR-PE. Basically, when the value of k increases ($k = [5, 30]$), recall improves while precision drops with a slow pace. When k gets even larger ($k = [30, 45]$), the improvement, nevertheless, is less obvious. This hints that setting the k value in between 30 and 45 can already offer very competitive performance to the exhaustive search. An interesting note is that filtering always offers better precision, indicating that the scheme is capable of pruning false matches not recognized by SR-PE. On the other hand, the value of recall degrades compared to the exhaustive search. This is most probably due to the quantization loss when performing clustering, which is still an open issue for further research.

The filtering scheme indeed significantly reduces the number of keyframe pairs to match and evaluate. As shown in Table V, there are 31,598 pairs (compared to 24,538,515 pairs in exhaustive search), achieving a speed-up ratio of approximately 750 times. When the value of k approaches to 45, the speed-up ratio is still as high as 95 times.

Table VI details the time spent for the fast detection of near-duplicates. The indexing and detection costs required to process the 24,538,515 of candidate pairs in CityU dataset are listed. For online detection, VK retrieval, which returns top-45 candidates, only needs 3 min to compare all images in the candidate pool. Keypoint matching, on the other hand, consumes most of the time by spending 6 h to process the 2% candidates retained by VK. SR-PE only needs few more minutes to further complete the task. Totally the online ND detection needs no more than 7 h. Compared to the exhaustive matching which can take more than one month, the algorithm with VK filtering is significantly efficient.

E. Discussion: Challenge and Difficulty

In our experiments, SR-PE has demonstrated better performance than other approaches. When coupling with VK filtering, ND detection is significantly sped up by the need to examine only a small portion of ND candidates. SR-PE is indeed a generalization of PE by evaluating candidate ND regions through the estimation of their 2-D transformations. There are two issues

TABLE VI
TOTAL TIME SPENT FOR ND DETECTION IN CITYU DATASET

	Step	*Speed*
Offline indexing	Keypoint extraction	10.9 hour
	Building inverted file	2.38 hour
Online detection	VK retrieval	3 min
	Keypoint matching	6.46 hour
	SR-PE prediction	4.25 min

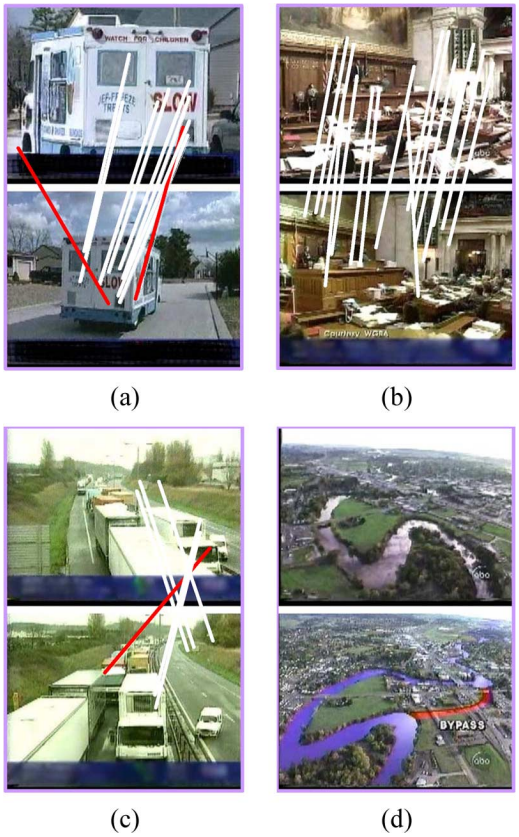


Fig. 10. ND pairs with different variations due to viewpoint changes: (a), (b) correctly detected by SR-PE; (c), (d) failure examples. (a) Object view variation; (b) scene view variation; (c) depth variation; (d) no matching line.

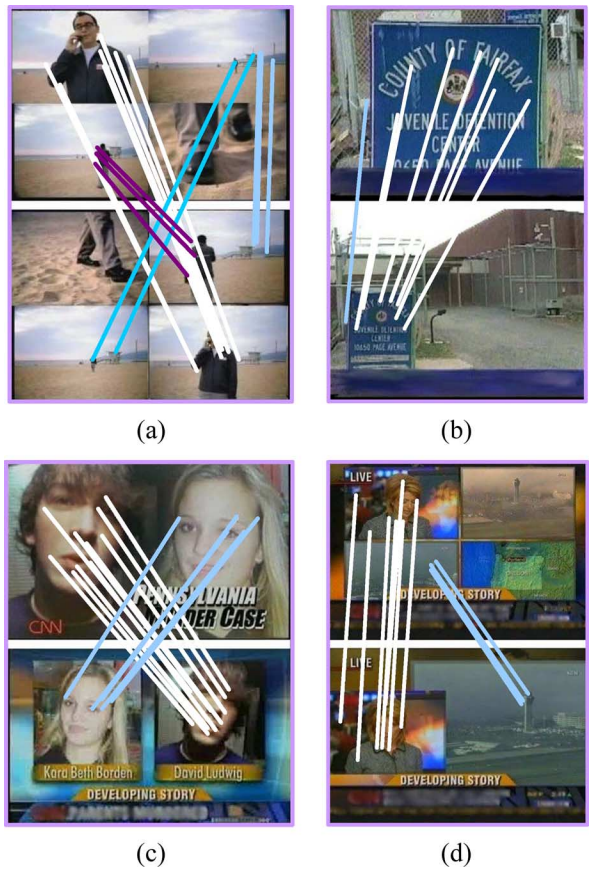


Fig. 11. Matched ND regions by SR-PE after clustering of matching lines. Lines from different ND regions are marked with different colors. (a) Scene and object duplicate; (b) object duplicate; (c) multiple-object duplicate; (d) multiple-scene duplicate.

worth to discuss: 1) the effectiveness of SR-PE when viewpoint changes in which 3-D transformation is normally involved, and 2) the extent to which SR-PE can be used to infer higher level semantics by localizing object and background duplicates.

Recognizing viewpoint change has long been known as a challenging problem particularly when using only 2-D transformation for 3-D estimation. To study this problem, a set of 40 ND image pairs involving viewpoint change is identified from CityU dataset. An interesting observation is that among the 40 pairs, only 24 pairs have keypoints matched to each other. This indeed implies two fundamental challenges when viewpoint changes: the difficulty in locating the right set of keypoints for matching and the robustness of SIFT-based descriptor in depicting keypoint features. Among the 24 ND image pairs with matching lines, SR-PE successfully detects 18 pairs. Other approaches such as RS-CT and CT are able to detect 14 and 12 pairs, respectively. This result somehow shows the difficulty of detecting near-duplicates when viewpoint changes, while also indicating the capability of SR-PE over other approaches in detecting this type of ND pairs. Fig. 10 shows few success and failure examples by SR-PE. Notice that these ND pairs not only differ in terms of viewpoint but also the changes such as scaling and illumination effects. Fig. 10(c) shows a typical example when most approaches including SR-PE will fail. Basically, when the scene depth varies a lot, the deformation caused by camera projection due to depth variation can introduce inaccuracy to 3-D estimation, particularly if the estimation is relied only on 2-D transformation. In our experiments, employing only 2-D scale and rotation still performs reasonably, as long as the scene depth does not vary a lot, as shown in Fig. 10(a) and (b). In addition, the viewpoint change should not cause problem to the localization of keypoints for robust matching. The example shown in Fig. 10(d), where viewpoint change causes problem of matching any available keypoints, is a typical failure example for keypoint-based approaches.

The second issue is regarding the capability of SR-PE in localizing the near-duplicate regions. Fig. 11 shows few examples where SR-PE is able to locate the duplicate objects and background in the images after the clustering of matching lines using 2-D transformation. This, indeed, indicates the potential of SR-PE in performing video mining related tasks such as hyperlinking content information across videos for search re-ranking. While the result is encouraging in our experiments, segmenting duplicate objects and background scenes for semantic representation or recognition remains difficult. Other

visual cues, in addition to the matched keypoints, are required for the accurate extraction of duplicate regions.

VII. Conclusion

We have presented our works for near-duplicate detection. In particular, the qualitative evaluation of keypoint matching patterns, when there are more than one ND regions with different transformations, are thoroughly addressed. Compared to PE, SR-PE is capable of evaluating complex patterns composed of ND regions under unknown scale and rotation changes. Empirical results indicate that, with the use of PSIFT descriptor, SR-PE performs better than other existing measures. To demonstrate the practical detection of near-duplicates in large dataset, we have also proposed a hierarchical framework with BoW, keypoint matching and SR-PE. Empirically, the framework is able to speed up the detection as fast as hundred times compared to exhaustive evaluation.

Our works can be extended for two other interesting tasks. First, SR-PE could be directly applied to evaluate the variants of keypoint detectors and descriptors. Powered by entropy, SR-PE can be an excellent measure for characterizing the discriminative-ness of detectors and descriptors. Second, our framework for ND detection can be adopted for hyper-linking of near-duplicate web images and videos, which would greatly facilitate multimedia search on Web.

References

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

[2] 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. Vis.*, vol. 65, no. 1/2, pp. 43–72, 2005.

[3] J. Zhang, M. Marszatek, 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, 2007.

[4] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," *ACM Multimedia*, pp. 869–876, 2004.

[5] S. F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.-Q. Zhang, "Columbia University TRECVID-2005 video search and high-level feature extraction," presented at the TRECVID, 2005.

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

[7] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE. Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[8] 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.

[9] X. Wu, A.-G. Hauptman, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *ACM Multimedia*, 2007, pp. 218–227.

[10] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang, "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," *ACM Multimedia*, pp. 845–854, 2006.

[11] J.-H. Hsiao, C.-S. Chen, L.-F. Chien, and M.-S. Chen, "A new approach to image copy detection based on extended feature sets," *IEEE. Trans. Image Process.*, vol. 16, no. 8, pp. 2069–2079, Aug. 2007.

[12] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos," *IEEE Signal Process. Mag.*, pp. 59–68, Mar. 2006.

[13] M. Shneider and S.-F. Chang, "A robust content based digital signature for image authentication," in *Proc. Int. Conf. Image Processing*, 1996, vol. 3, pp. 227–230.

[14] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," *ACM Multimedia*, pp. 877–884, 2004.

[15] 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.

[16] K. Mikoljczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, pp. 63–86.

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable resgions," in *Proc. Brit. Machine Vision Conf.*, 2002, pp. 384–396.

[18] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 506–513.

[19] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa, "Robust voting algorithm based on labels of behavior for video copy detection," *ACM Multimedia*, pp. 835–844, 2006.

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

[21] M. Heritier, S. Foucher, and L. Gagnon, "Key-plcaces detection and clustering in movies using latent aspects," in *Proc. Int. Conf. Image Processing*, 2007, pp. 225–228.

[22] M. Heritier, L. Gagnon, and S. Foucher, "Key-plcaces clustering of full-length film key-frames using latent aspects modeling over SIFT matches," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[23] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[24] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: A texture classification example," presented at the Int. Conf. Computer Vision, 2003.

[25] [Online]. Available: http://www.ee.columbia.edu/ln/dvmm

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

[27] TREC Video Retrieval Evaluation (TRECVID) [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[28] K. Mikoljczyk and C. Schmid, "An affine invariant interest point detector," presented at the Eur. Conf. Computer Vision, 2002.

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

Wan-Lei Zhao received the B.Eng. and M.S. degrees from the Department of Computer Science and Engineering, Yunnan University, China, in 2002 and 2006, respectively. He is currently pursuing the Ph.D. degree in the Department of Computer Science, City University of Hong Kong.

He was with the Software Institute, Chinese Academy of Sciences, from 2003 to 2004 as an exchange student. His research interests include video computing and machine learning.

Chong-Wah Ngo (M'02) received the B.Sc. and M.Sc. degrees in computer engineering from the Nanyang Technological University of Singapore and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology in 2000.

Before joining the City University of Hong Kong in 2002, he was with the Beckman Institute, University of Illinois at Urbana-Champaign, Urbana. He was also a Visiting Researcher with Microsoft Research Asia in 2002. His research interests include video computing and multimedia information retrieval.