Near-Duplicate Keyframe Retrieval by Nonrigid Image Matching

Jianke Zhu
CSE Department, Chinese University of Hong Kong, Hong Kong
jkzhu@cse.cuhk.edu.hk

Steven C.H. Hoi
Computer Eng. School, Nanyang Technological University, Singapore
chhoi@ntu.edu.sg

Michael R. Lyu
CSE Department, Chinese University of Hong Kong, Hong Kong
lyu@cse.cuhk.edu.hk

Shuicheng Yan
ECE Department, National University of Singapore, Singapore
eleyans@nus.edu.sg

ABSTRACT

Near-duplicate image retrieval plays an important role in many real-world multimedia applications. Most previous approaches have some limitations. For example, conventional appearance-based methods may suffer from the illumination variations and occlusion issue, and local feature correspondence-based methods often do not consider local deformations and the spatial coherence between two point sets. In this paper, we propose a novel and effective Nonrigid Image Matching (NIM) approach to tackle the task of near-duplicate keyframe retrieval from real-world video corpora. In contrast to previous approaches, the NIM technique can recover an explicit mapping between two near-duplicate images with a few deformation parameters and find out the correct correspondences from noisy data effectively. To make our technique applicable to large-scale applications, we suggest an effective multi-level ranking scheme that filters out the irrelevant results in a coarse-to-fine manner. In our ranking scheme, to overcome the extremely small training size challenge, we employ a semi-supervised learning method for improving the performance using unlabeled data. To evaluate the effectiveness of our solution, we have conducted extensive experiments on two benchmark testbeds extracted from the TRECVID2003 and TRECVID2004 corpora. The promising results show that our proposed method is more effective than other state-of-the-art approaches.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithm, Performance, Experimentations

Keywords

Near-Duplicate Keyframe, Image Copy Detection, Nonrigid Image Matching, Semi-supervised Learning

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

1. INTRODUCTION

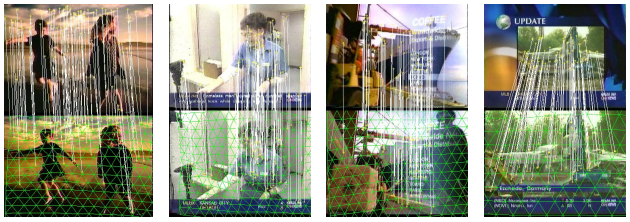
Near-Duplicate Keyframes (NDK) refer to the pairs of keyframes in a video corpus, for which the two keyframes of a pair are closely similar to each other apart from minor differences due to the variations of capturing conditions, rendering conditions, or editing operations [30, 32]. NDK detection and retrieval techniques are beneficial for many real applications, such as news video search [23] and copyright infringement detection [13, 20]. NDK retrieval is a challenging research problem due to some well-known factors. One is that videos from different sources may be captured by devices with different hardwares under a variety of illumination conditions. Moreover, video editing often produces extra photometric and geometric transformations and occludes the original video by adding captions. Figure 1 shows some examples of pairs of duplicate keyframes extracted from the TRECVID2003 video corpus.

In the past years, there has been a surge of research attention on this topic in the multimedia community[13, 20, 26, 27, 30, 32]. Some conventional methods extend content-based image retrieval (CBIR) techniques for the NDK detection and retrieval task; these often employ global features extracted from the whole image, such as color moment and color histogram [20, 30]. Although these methods are usually very efficient for finding identical copies, they may not be very accurate for real NDKs as they often fail to address the variations of lighting changes, viewpoint changes, and occlusions.

Alternatively, some recent approaches using local feature point correspondences can deal with the illumination variations and geometric transformations by exploring the recent advances in local feature descriptors [16]. These approaches often incur heavy computational cost in feature matching. Nevertheless, some efficient solutions have been proposed. For example, Ke et al. [13] proposed an efficient method using PCA-SIFT and locality-sensitive hashing indexing. However, their method often makes a *rigid* projective geometry assumption, which may suffer from some outlier matches due to lens changes and small object movements. Zhang and Chang [30] presented a stochastic Attributed Relational Graph (ARG) matching framework, which involves a computationally intensive process of stochastic belief propagation. Zhao et al. [32] proposed a one-to-one symmetric (OOS) matching method, which applies a local smoothing constraint to remove the outlier matches. In [17], Pattern Entropy (PE) is employed as similarity measure for OOS method. Similar to other *bipartite graph matching* methods, the OOS method considers only pairwise matches

and fails to explore the *spatial coherence* between the two sets of interest points in two NDKs. As shown in Figure 1, illumination variations, occlusions and zooming lead to large PE, in which $PE \leq 0.5$ is considered as NDK pair [17].

In contrast to previous approaches employing either rigid projective models or bipartite graph matching, in this paper, we propose a novel **Nonrigid Image Matching** (NIM) method for near-duplicate keyframe retrieval. Unlike the previous approaches, we assume that there may exist nonrigid transformations between the two NDKs. The key to solving the NIM problem is based on an iterative coarse-to-fine optimization scheme to reject the outliers, which takes advantage of a closed-form solution for a given set of correspondences. Since our method takes consideration of local deformations, it often obtains more inlier matches than regular rigid projective models and the OOS graph matching method, this characteristic plays a critical role in duplicate similarity matching. Figure 1 shows some examples along with the total numbers of inlier matches found by three different methods on the same set of extracted SIFT features [15].



(a) $19/\frac{52}{0.73}/92$ (b) $26/\frac{58}{0.76}/107$ (c) $15/\frac{75}{0.72}/104$ (d) $56/\frac{76}{0.68}/148$

Figure 1: Some near-duplicate keyframes examples. The caption of each subfigure shows the total number of inlier matches with each of the three methods: projective geometry, OOS-SIFT method (PE is below the number of inliers), and our NIM method. Since $PE > 0.5$, OOS-SIFT method failed in (a-d).

Compared to the previous approaches, the proposed NIM method not only delivers better retrieval performance, but also enjoys some other salient merits. For example, our method is able to find the exact matching region between two NPKs, which is often not obtained by conventional methods. This attractive feature is important for part-based or sub-image detection and retrieval. In addition, our method is rather efficient, processing about ten pairs of keyframes per second with a regular PC with moderate configuration. To further accelerate our technique for large-scale applications, we suggest a Multi-Level Ranking (MLR) framework for efficient NDK retrieval, which integrates three different ranking components in a unified solution: nearest neighbor ranking, semi-supervised ranking, and NIM-based ranking.

In summary, this paper includes three main contributions. First of all, we propose a novel **Nonrigid Image Matching** technique for NDK detection and retrieval, which is significantly different from the conventional approaches. Our technique overcomes some limitations with the existing approaches and hence offers better performance for solving the NDK detection and retrieval tasks. Secondly, to enable the proposed technique applicable to large-scale applications, we suggest a **Multi-Level Ranking** framework that can effectively filter out irrelevant results so as to significantly reduce the sample size for the NIM comparisons. Although this is not the first use of the MLR approach by multimedia researchers [10, 11], our contribution is to validate its effectiveness at improving the NIM scheme in the NDK retrieval tasks. The third major contribution is to employ a **Semi-Supervised Ranking** (SSR) method by a *Semi-Supervised Support Vector Machine* (S^3VM) for improving the NDK learning task, which often has extremely few labeled data. The SSR method effectively improves the filtering performance of traditional supervised learning approaches by taking advantage of unlabeled data information.

The rest of this paper is organized as follows. Section 2 reviews some existing approaches for NDK detection and retrieval. Section 3 proposes the nonrigid image matching method for detecting NDK with local feature correspondences. Section 4 presents a multi-level ranking scheme together with a semi-supervised SVM method for NDK retrieval. Section 5 provides our experimental results and the details of our experimental implementation. Section 6 sets out our conclusions.

2. RELATED WORK

There are numerous research efforts on near-duplicate image/keyframe detection and retrieval in the multimedia community [13, 20, 25, 27, 30]. In general, most of the existing approaches can be roughly divided into two categories: *appearance-based methods* and *local feature-based methods*.

The appearance-based methods often measure the similarity between two keyframes based on the extracted global visual features, such as color histogram [30] and color moments [31]. These methods are advantageous for their high efficiency since keyframes are often compactly represented in the vector space and thus can be solved efficiently by adapting conventional CBIR methods and mature data indexing techniques [20]. But they are often not very robust to illumination changes, partial occlusions, and geometric transformations.

On the other hand, the local feature-based methods detect local keypoints in two keyframes and measure their similarity by counting the number of correct correspondences between two keypoint sets. Keypoints are the salient regions detected over image scales and their descriptors are often invariant to certain transformations and variations. They overcome the limitations of the global appearance-based methods, and thus often achieve better performance [13, 32]. But they may incur a heavy computational cost for the matching of two keypoint sets, which may contain more than one thousand of keypoints.

Recently, local feature-based methods have been actively studied. Sivic et al. [22] employed the local keypoints approach for object matching and retrieval in movies. Ke et al. [13] employed the compact PCA-SIFT feature and speeded up the search of nearest keypoints with the locality sensitive hashing technique for duplicate image detection and retrieval. Zhao et al. [32] proposed an OOS matching approach to NDK detection and reported state-of-the-art performance. The key of the OOS method is to eliminate noisy outliers during the one-to-one bipartite graph matching process. Most of these methods fall in the same category of point-to-point bipartite graph matching.

The NIM technique proposed in this paper goes beyond conventional point-to-point bipartite graph matching methods. In contrast to existing techniques, our method is able to recover the explicit mapping between two near-duplicate

keyframes with nonrigid transformation models and can effectively find the correct correspondences from noisy data. Though similar techniques are actively being studied for tracking in computer vision and graphics [35, 36], to the best of our knowledge, we are the first to study it comprehensively for NDK retrieval tasks.

3. NONRIGID IMAGE MATCHING

In this section, we present the nonrigid image matching approach to near-duplicate keyframe detection. We first give our formulation of the nonrigid image matching problem, and then solve it by a coarse-to-fine optimization technique.

3.1 Formulation

Instead of assuming an affine transformation or projective projection as in the conventional methods, we employ the nonrigid mapping relation between the NDKs. Therefore, the proposed method can tackle not only geometric transformations and viewpoint changes, but also small object movements. The *Nonrigid Image Matching* refers to the problem of recovering the explicit mapping between the two images with a few deformation parameters and finding out the correct correspondences from noisy data simultaneously. It has been successfully applied to real-time nonrigid surface tracking in computer vision [19, 35, 36]. Unlike the nonrigid image registration, the NIM method is fully automatic and does not require manual initialization.

The key idea of NIM is to recover the local deformations from the salient feature correspondences between the two images and to reject the outlier matches simultaneously. Therefore, we can simply choose the total number of inlier matches τ as a confidence measure to judge whether the two keyframes are near-duplicate or not. Specifically, given a set of correspondences $\mathcal{M}$ between the model and the input image built through a local feature matching algorithm, we try to estimate the nonrigid mapping from these observations. Therefore, a pair of matched points is represented in the form of $\mathbf{m} = \{\mathbf{m}_0, \mathbf{m}_1\} \in \mathcal{M}$, where $\mathbf{m}_0$ is defined as the 2D coordinates of a feature point in the training image and $\mathbf{m}_1 = (u, v)$ is the coordinates of its match in the input image. We represent the query keyframe as a deformation grid, which is explicitly represented by triangulated meshes with N hexagonally connected vertices. The vertices' coordinates are formed into a shape vector $\mathbf{s} = (\mathbf{u} \quad \mathbf{v})^\top$, where $\mathbf{u} \in R^N$ and $\mathbf{v} \in R^N$ are the vectors of the coordinates of mesh vertices. Therefore, $\mathbf{s}$ is the variable to be estimated from the 2D correspondences.

We commence by assuming that a point $\mathbf{m}$ lies in a triangle whose three vertices' coordinates are (u_i, v_i),(u_j, v_j) and (u_k, v_k) respectively, and $\{i, j, k\} \subset [1, N]$ is the index of each vertex. The piecewise affine transformation is used to map the image points inside the corresponding triangle into the vertices in the mesh. Thus, the mapping function $T_S(\mathbf{m})$ is defined as below:

$$T_{\mathbf{s}}(\mathbf{m}) = \begin{bmatrix} u_i & u_j & u_k \\ v_i & v_j & v_k \end{bmatrix} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \end{bmatrix}^\top \tag{1}$$

where (ξ_1, ξ_2, ξ_3) are the barycentric coordinates for the point $\mathbf{m}$, and $\xi_1 + \xi_2 + \xi_3 = 1$.

Then, the correspondence error $E_c(\mathbf{s})$ is defined as the sum of the weighted square error residuals for the matched points, which can be formulated as follows:

$$E_c(\mathbf{s}) = \sum_{\mathbf{m} \in \mathcal{M}} \omega_{\mathbf{m}} \mathcal{V}(\delta, \sigma) \tag{2}$$

where $\mathcal{V}(\delta, \sigma)$ is a robust estimator with compact support size σ, and $\omega_{\mathbf{m}} \in [0, 1]$ is a weight linked with each correspondence. Moreover, δ is the residual error, which is defined as $\delta = \mathbf{m}_1 - T_{\mathbf{s}}(\mathbf{m}_0)$.

The robust estimator function $\mathcal{V}(\delta, \sigma)$ that assesses a fixed penalty for residuals larger than a threshold σ is employed in the present work; this approach is relatively insensitive to outliers [4]:

$$\mathcal{V}(\delta, \sigma) = \begin{cases} \frac{\|\delta\|}{\sigma^\nu}, & \mathcal{M}_1 = \{\mathbf{m} | \quad \|\delta\| \le \sigma^2\} \\ \sigma^{2-\nu}, & \mathcal{M}_2 = \overline{\mathcal{M}_1} \end{cases} \tag{3}$$

where the set $\mathcal{M}_1$ contains the inlier matches, and $\mathcal{M}_2$ is the set of the outliers. In addition, the order ν determines the scale of the residual. The greatest number of correspondences is included when the support σ is large. Conversely, as σ decreases, the robust estimator becomes narrower and more selective.

In general, the NIM problem approximates a 2D mesh with N vertices from the keypoint correspondences, which is usually ill-posed. One effective way to attack this problem is to introduce regularization, which preserves the regularity of a deformable mesh and constrains the searching space. The following object function is widely used in deformable surface fitting [12, 19] for energy minimization:

$$E(\mathbf{s}) = E_c(\mathbf{s}) + \lambda_r E_r(\mathbf{s}) \tag{4}$$

where $E_r(\mathbf{s})$ is the regularization term that represents the deformation energy, and λ_r is a regularization coefficient. The regularization term E_r in the above equation, also known as 'internal force' in Snakes [12], is composed of the sum of the squared second-order derivatives of the mesh vertex coordinates. As the mesh is regular, $E_r(\mathbf{s})$ can be formulated through a finite difference:

$$E_r(\mathbf{s}) = \mathbf{s}^\top \begin{bmatrix} \mathcal{K} & 0 \\ 0 & \mathcal{K} \end{bmatrix} \mathbf{s} \tag{5}$$

where $\mathcal{K}$ is a sparse and banded matrix which is determined by the structure of the explicit mesh model [8, 19].

3.2 Optimization

As the robust estimator function in Eqn. 3 is not convex, this leads to a hard combinational optimization problem for the associated penalty function approximation. To tackle this problem, we employ a progressive finite Newton optimization method [35, 36]. Given a set of inlier matches $\mathcal{M}_1$, the solution for the optimization problem in Eqn. 4 can be obtained by solving the following two linear equations via LU decomposition:

$$\mathbf{u} = (\lambda_r \mathcal{K} + A)^{-1} \mathbf{b}_u \tag{6}$$

$$\mathbf{v} = (\lambda_r \mathcal{K} + A)^{-1} \mathbf{b}_v \tag{7}$$

where $A \in R^{N \times N}$ is equal to

$$A = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^\nu} \mathbf{t}\mathbf{t}^\top$$

and the vector $\mathbf{b}_u \in R^N$ and $\mathbf{b}_v \in R^N$ are defined as below:

$$\mathbf{b}_u = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^\nu} u\mathbf{t} \quad \text{and} \quad \mathbf{b}_v = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^\nu} v\mathbf{t}$$

where $\mathbf{t} \in R^N$ containing the barycentric coordinates is defined as below:

$$\mathbf{t}_i = \xi_1 \quad \mathbf{t}_j = \xi_2 \quad \mathbf{t}_k = \xi_3$$

while the remaining elements in the vector $\mathbf{t}$ are all set to zero. It can be observed that the overall complexity of the NIM method is that of a single Newton step, which is determined by the total number of mesh vertices N.

Obviously, we can directly compute $\mathbf{s}$ by the above closed-form solution if the correspondences set $\mathcal{M}$ contains no outliers. However, the incorrect matches cannot be avoided in the first stage of the matching process where only local image descriptors are compared. Therefore, a coarse-to-fine optimization scheme is introduced to reject the outliers gradually, which progressively decays the support σ of the robust estimator $\mathcal{V}(\delta, \sigma)$ at a constant rate η. For each value of σ, the object function E is minimized through the finite Newton step and the result is employed as the initial state for the next minimization. The optimization procedure stops when σ reaches a value close to the expected precision, which is usually one or two pixels. Thus, the whole optimization problem can be solved within a finite number of steps. As the derivatives of $\mathcal{V}(\delta, \sigma)$ are inversely proportional to the support σ, the regularization coefficient λ_r is kept constant during the optimization.

Before starting the optimization, we need to select the initial active set. One strategy is to set the initial value of σ to a sufficiently large value in order to select most of the correspondences into the initial active set and to avoid getting stuck at local minima. This method may need a few steps to compensate for the errors generated by the variations in object positions between the images. Alternatively, we can select the active set through a modified RANSAC [6, 7] approach by taking advantage of our closed-form solution. Note that it is usually hard to directly apply the robust estimator to a system with a large number of free variables. To reduce the total number of RANSAC trials, we draw from progressively larger sets of top-ranked correspondences with the highest similarities. In the experiments, the sampling process stopped within five trials. Since the result of RANSAC is usually quite close to the solution, the initial value of σ can be relatively small. Thus, the proposed progressive scheme requires fewer steps.

3.3 Case Studies: Detecting Various NDKs

To illustrate how the proposed NIM technique can effectively detect various NDKs appearing in news video domains, we show part of our detection results to demonstrate the advantages of our technique.

Figure 2 shows some examples of our successful detection results for various NDKs. All results on the duplicate pairs from Columbia's TRECVID2003 dataset can be found at [2]. In particular, the proposed NIM technique can effectively detect a variety of NDKs including, but not limited to, the following cases:

- **Viewpoint change.** This is very common for the shots extracted from news video sequences.
- **Object movement.** This is due to the relative movements caused by the camera or some objects.
- **Lens change.** This case is caused by the changes of camera lens, such as zooming in or zooming out.
- **Subimage duplicate.** Such duplicates could be caused either by lens changes or some editing effects.
- **Small regional change.** These duplicates only have small regional differences. They are often captured in the same scenario with slight changes.
- **Partial occlusion.** This case arises from the added captions or text descriptions in the videos.

4. MULTI-LEVEL NDK RETRIEVAL

4.1 Framework Overview

Although the proposed NIM is efficient for matching two images in comparison with conventional local feature matching techniques [27, 32], directly applying NIM to large-scale applications could still be computationally intensive. To improve the efficiency and scalability of our solution, we employ a Multi-Level Ranking (MLR) framework for efficiently tackling the NDK retrieval task. This strategy has been widely used, which is also shown to be successful in multimedia retrieval [10, 11]. In particular, our multi-level ranking scheme integrates three different ranking components:

- **Nearest Neighbor Ranking (NNR).** This is to rank the keyframes with simple nearest neighbor search.
- **Semi-Supervised Ranking (SSR).** This is to rank the keyframes with a semi-supervised ranking method.
- **Nonrigid Image Matching (NIM).** This is to rank the keyframes by applying the proposed NIM method.

The first two ranking components are based on global features for efficiently filtering out the irrelevant results, and the last component provides a fine re-ranking based on the local features. Figure 3 shows the proposed MLR framework, which attacks the NDK retrieval task in a coarse-to-fine ranking manner. This makes the proposed NIM solution applicable to large-scale real applications.

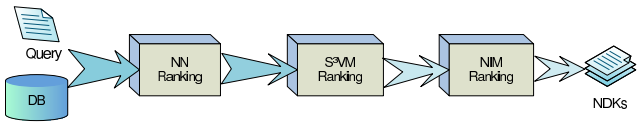
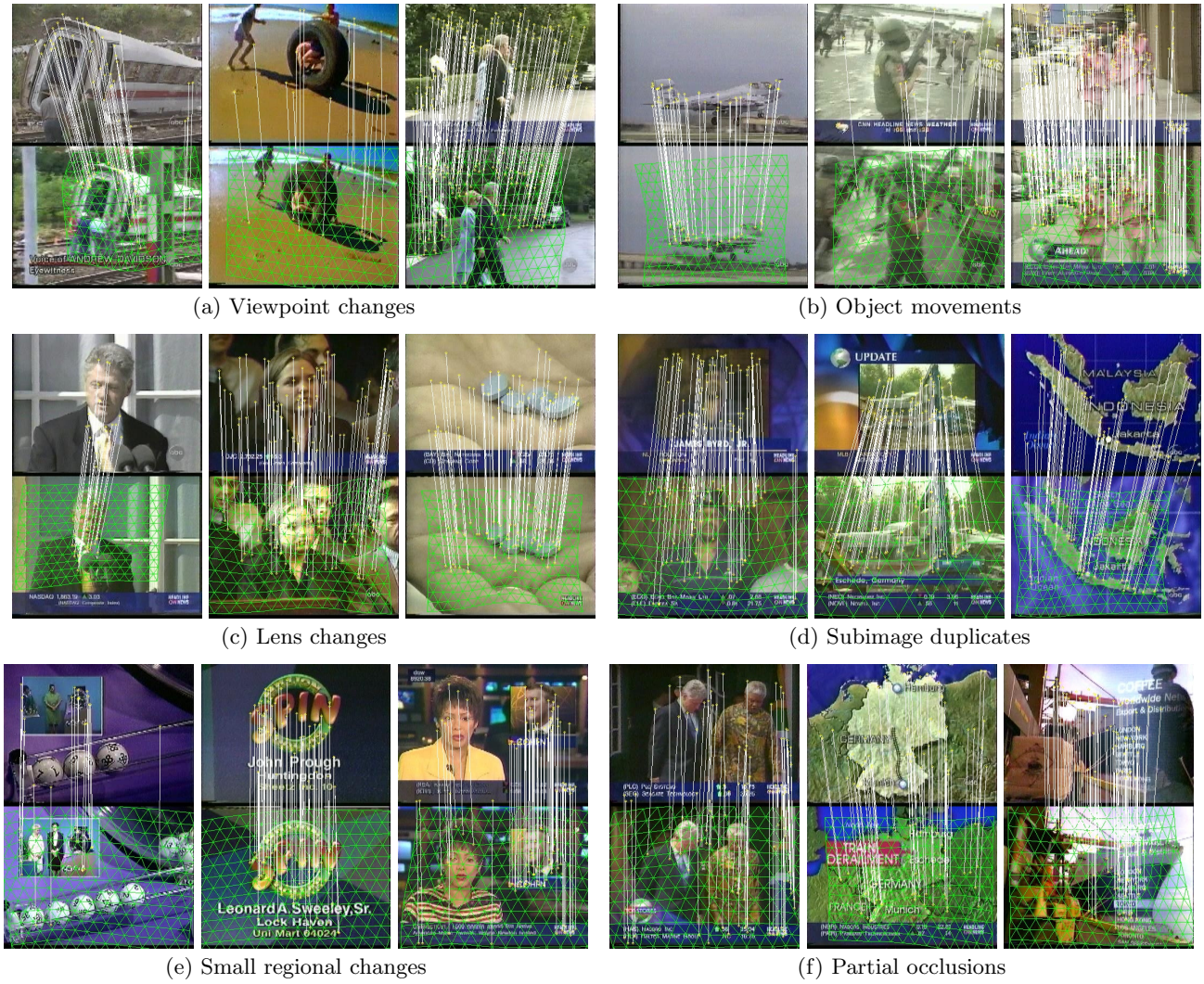


Figure 3: A multi-level ranking framework.

4.2 Formulation as a Machine Learning Task

The NDK retrieval problem can be formulated as a machine learning task with a query set of labeled image examples $\mathcal{Q} = \{(\mathbf{x}_1, +1), \ldots, (\mathbf{x}_l, +1)\}$ and a gallery set of unlabeled image examples $\mathcal{G} = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}\}$, where each image example $\mathbf{x}_i \in R^d$ is represented in a d-dimensional feature space. The goal of the learning task is to find the relevant near-duplicate examples from $\mathcal{G}$ that are closest to being exact duplicates of examples in $\mathcal{Q}$.

The learning task is tough on account of two difficulties. One is that there is no negative examples available, as only a query set $\mathcal{Q}$ will be provided in the retrieval task. The other is the small sample learning issue: Very few labeled examples will be provided in the retrieval task. To overcome the first difficulty, we adopt the idea of pseudo-negative examples used in previous multimedia retrieval approaches [29]. Specifically, we can conduct a query-by-example retrieval



(a) Viewpoint changes (b) Object movements

(c) Lens changes (d) Subimage duplicates

(e) Small regional changes (f) Partial occlusions

Figure 2: Examples of our detection results on various near-duplicate keyframe cases.

for ranking the unlabeled data in $\mathcal{G}$ based on their distances from the examples in the query set. Then we select a short list of most dissimilar examples as the negative examples based on the Nearest Neighbor ranking results.

To this end, with both positive and negative examples, we can formulate the learning task as a general binary classification task, which can then be solved by existing classification techniques. In our approach, we apply Support Vector Machines (SVM) for the learning task. SVM is a well-known and state-of-the-art learning technique [24], which we briefly review here. SVM is used for learning an optimal hyperplane with maximal margin, and can learn nonlinear decision boundaries by exploiting powerful kernel tricks. SVM can be generally formulated in a regularization framework:

$$\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} \max(0, 1 - y_i f(\mathbf{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \qquad (8)$$

where f is the hyperplane function $f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$, k is some kernel function, and $\mathcal{H}_K$ is the associated reproducing kernel Hilbert space.

While SVM can be applied for solving the learning task, its performance may be poor when there are very limited number of labeled examples. This is a critical issue of an NDK retrieval since only extremely few positive examples will be provided. To overcome the second difficulty, we next introduce a semi-supervised learning technique for exploring both labeled and unlabeled data for the retrieval tasks.

4.3 Semi-supervised Support Vector Machine

To overcome the challenge of small sample learning, we suggest a semi-supervised retrieval (SSR) approach to attack the learning task via a semi-supervised SVM technique. Semi-supervised learning has been extensively studied in recent years, and numerous approaches have been proposed to exploit it [28, 33, 34]. In this paper, we employ a unified kernel learning approach for semi-supervised SVM. The key idea is to first learn a data-dependent kernel from the unlabeled data, and then apply the learned kernel to train a supervised SVM based on the regularization learning framework. In our approach, we adopt the kernel deformation principle for learning a data-dependent kernel from unlabeled data [21].

The main idea of kernel deformation is to first estimate the geometry of the underlying marginal distribution from

both labeled and unlabeled data, and then derive a data-dependent kernel by incorporating estimated geometry [21]. Let $\mathcal{H}$ denote the original Hilbert space reproduced by kernel function $k(\cdot,\cdot)$, and $\widetilde{\mathcal{H}}$ denote the deformed Hilbert space. In [21], the authors assume the following relationship between the two Hilbert spaces:

$$< f, g >_{\widetilde{\mathcal{H}}} = < f, g >_{\mathcal{H}} + \mathbf{f}^\top M \mathbf{g}$$

where $f(\cdot)$ and $g(\cdot)$ are two functions, $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_1))$ evaluates the function $f(\cdot)$ for both labeled and unlabeled data, and M is the distance metric that captures the geometric relationship among all the data points. The deformation term $\mathbf{f}^\top M \mathbf{g}$ is introduced to assess the relationship between the functions $f(\cdot)$ and $g(\cdot)$ based on the observed data. Given an input kernel k, the explicit form of the new kernel function $\tilde{k}$ can be derived as below:

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) + \kappa_{\mathbf{y}}^\top \mathbf{d}(\mathbf{x})$$

where $\kappa_{\mathbf{y}} = (k(\mathbf{x}_1, y), \ldots, k(\mathbf{x}_n, y))^\top$. The coefficients vector $\mathbf{d}(\mathbf{x})$ can be computed by: $\mathbf{d}(\mathbf{x}) = -(I + MK)^{-1} M \kappa_{\mathbf{x}}$, where $K = [k(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$ is the original kernel matrix for all the data, and $\kappa_{\mathbf{x}} = (k(\mathbf{x}_1, z), \ldots, k(\mathbf{x}_n, z))^\top$. To capture the underlying geometry of the data, a common approach is to define M as a function of graph Laplacian L, for example, $M = L^p$ where p is an integer. A graph Laplacian is defined as $L = \mathrm{diag}(S\mathbf{1}) - S$, where $S \in R^{n \times n}$ is a similarity matrix and each element $S_{i,j}$ is calculated by an RBF function $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / \varsigma^2)$. ς denotes the kernel width for a graph Laplacian. $\mathbf{1}$ denotes a vector with all one elements. Consequently, the new kernel k can be formulated as follows:

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) - \kappa_{\mathbf{y}}^\top (I + MK)^{-1} M \kappa_{\mathbf{x}} \quad (9)$$

Hence, replacing the kernel k in Eqn. 8 by the kernel $\tilde{k}$ in Eqn. 9, we can train the semi-supervised SVM classifier.

5. EXPERIMENTS

In this section, we report our empirical study of the proposed techniques for NDK retrieval. Two key techniques will be evaluated comprehensively in our experiments. The first experiment is to examine the effectiveness of the *Multi-Level Ranking* scheme for filtering out the irrelevant results. In particular, we would like to examine whether the semi-supervised ranking method using S^3VM is more effective than the conventional ranking approaches. The second and more important experiment is to evaluate the performance of the proposed NIM technique for NDK retrieval in comparison with some state-of-the-art approaches. In the following experiments, we mainly report quantitative evaluations.

5.1 Experimental Testbeds and Setup

To conduct comprehensive evaluations, we employ two benchmark datasets for NDK retrieval as our experimental testbeds. One is the widely used Columbia's TRECVID2003 dataset [30], which consists of 600 keyframes with 150 near duplicate image pairs and 300 non-duplicate images extracted from the TRECVID2003 corpus [30]. All the keyframes are with the same size, 352 × 264. The other is CityU's TRECVID2004 dataset [1] recently collected by Ngo et al. [17]. It contains 7,006 keyframes with 3,388 near-duplicate image pairs, which are selected from the TRECVID2004 video corpus. In the TRECVID2004 dataset, the near-duplicate image pairs involve a total of 1,953 keyframes, which is about 28% of the whole collection. Note that one keyframe may be associated with several near-duplicate pairs.

To make a fair comparison with the state-of-the-art approaches, we adopt the evaluation protocol used in [32]. Specifically, all NDK pairs are adopted as queries for performance evaluation. Each query set $\mathcal{Q}$ contains a single keyframe image; other remaining keyframes are regarded as the gallery set $\mathcal{G}$. For the retrieval task, each algorithm produces a list of relevant results by ranking the keyframes in the gallery set. To evaluate the retrieval performance, the average *cumulative accuracy* metric is adopted as a performance metric [32], in which the accuracy is measured by judging whether the retrieved keyframe is one of the corresponding pairwise duplicates in the ground truth query set. As a yardstick for assessing the performance, we compare our method with the recently proposed OOS matching algorithm [32], one state-of-the-art method for NDK detection and retrieval.

For the experimental setups, the kernel function used in both SVM and S^3VM is an RBF kernel with fixed width. Regarding the parameter settings, the penalty parameter C of SVMs is set to 10 (or $\gamma_A = 10^{-1}$) and the graph regularization parameter of S^3VM is set to $\gamma_I = 10^{-1}$.

All the experiments in this paper were carried out on a notebook computer with Intel Core-2 Duo 2.0GHz processor and 2GB RAM. All the proposed methods are implemented in Matlab, for which some routines are written in C code. The code can be downloaded from [2]

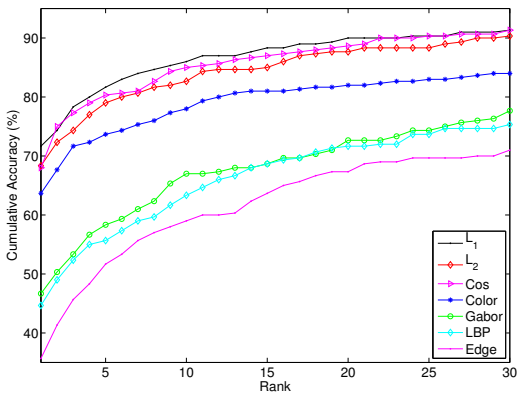
5.2 Feature Extraction

Feature extraction is a key step for NDK retrieval. In our experiments, we consider both global and local features. The two types of features have their advantages and disadvantages. We believe an appropriate fusion of them will compensate their shortcomings, and therefore improve the overall effectiveness and efficiency.

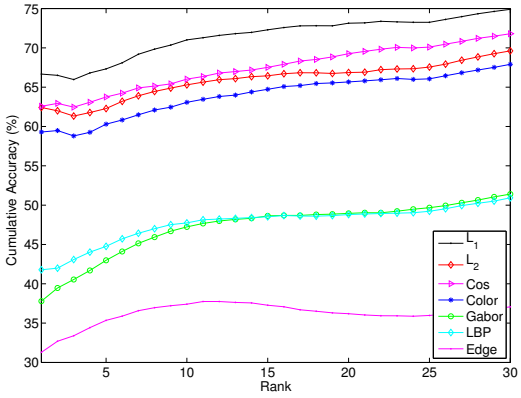
5.2.1 Global Feature Extraction

The global feature representation techniques have been extensively studied in image processing and CBIR community. A wide variety of global feature extraction techniques were proposed in the past decade. In this paper, we extract four kinds of effective global features:

- **Grid Color Moment.** We adopt the grid color moment to extract color features from keyframes. Specifically, an image is partitioned into 3 × 3 grids. For each grid, we extract three kinds of color moments: color mean, color variance and color skewness in each color channel (R, G, and B), respectively. Thus, an 81-dimensional grid color moment vector is adopted for color features.
- **Local Binary Pattern (LBP).** The local binary pattern [18] is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. In our experiment, a 59-dimensional LBP histogram vector is adopted.
- **Gabor Wavelets Texture.** To extract Gabor texture features, each image is first scaled to 64×64 pixels. The Gabor wavelet transform [14] is then applied on the scaled image with 5 levels and 8 orientations, which results in 40 subimages. For each subimage, 3 moments are calculated: mean, variance and skewness. Thus, a 120-dimensional vector is used for Gabor texture features.
- **Edge.** An edge orientation histogram is extracted for each image. We first convert an image into a gray image, and then employ a Canny edge detector [5] to obtain the edge map for computing the edge orientation histogram. The



(a) TRECVID2003 Dataset



(b) TRECVID2004 dataset

Figure 4: Cumulative accuracy of similarity measure and features using Nearest Neighbor Ranking on the TRECVID2003 dataset (600 keyframes) and the TRECVID2004 dataset (7006 keyframes).

edge orientation histogram is quantized into 36 bins of 10 degrees each. An additional bin is used to count the number of pixels without edge information. Hence, a 37-dimensional vector is used for shape features.

In total, a 297-dimensional vector is used to represent all the global features for each keyframe in the datasets.

5.2.2 *Local Feature Extraction*

Interest point detection and matching is a fundamental research problem in computer vision. Many effective approaches have been proposed in the literature. One of the most widely used methods is the SIFT [15], which computes a histogram of local oriented gradients around the interest point and stores the bins in a 128-dimensional vector. To improve the SIFT, Ke et al. [13] proposed an extended method by applying Principle Component Analysis [9] on the gradient image, which then yields a 36-dimensional descriptor that is more compact and faster for matching. However, the PCA-SIFT has been empirically shown to be less distinctive than the original SIFT in a comparative study [16], and is also slower than the original SIFT in the feature computation. Instead of using SIFT or PCA-SIFT, we adopt SURF [3], another emerging local feature descriptor to detect and extract local features, which takes advantage of fast feature extraction using integral images for image convolutions. Specifically, a 64-dimensional feature vector is used for representing each keypoint with SURF. Compared to the SIFT, it is more compact and hence reduces the computational cost for keypoint matching.

5.3 Experiment I: Ranking on Global Features

In this part, we evaluate the effectiveness of the proposed multi-level ranking scheme for filtering out the irrelevant keyframes by ranking on global features. We will first evaluate the retrieval performance of the global features with nearest neighbor ranking, and then evaluate the semi-supervised ranking approach based on S^{3}VM.

5.3.1 *Effectiveness of Global Features*

To examine how effective the global features are, we measure the retrieval performance of different distance measures with the global features on both datasets, as shown in Figure 4. From the results, we first observe that different distance metrics have different impacts on the retrieval results with the same global features. In particular, the L_1 norm outperforms both the L_2 norm and the cosine metric on both datasets, and the cosine similarity is slightly better than the L_2 norm. As a result, we employ the L_1 norm as the distance measure in all of the remaining experiments.

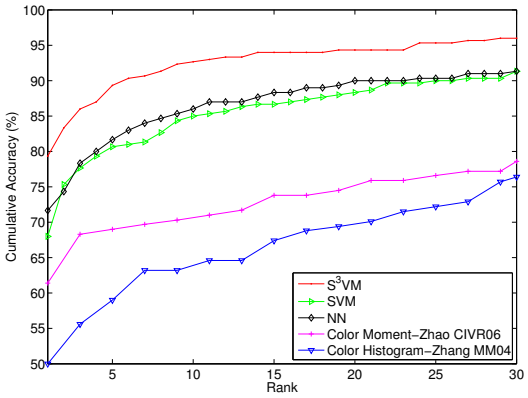
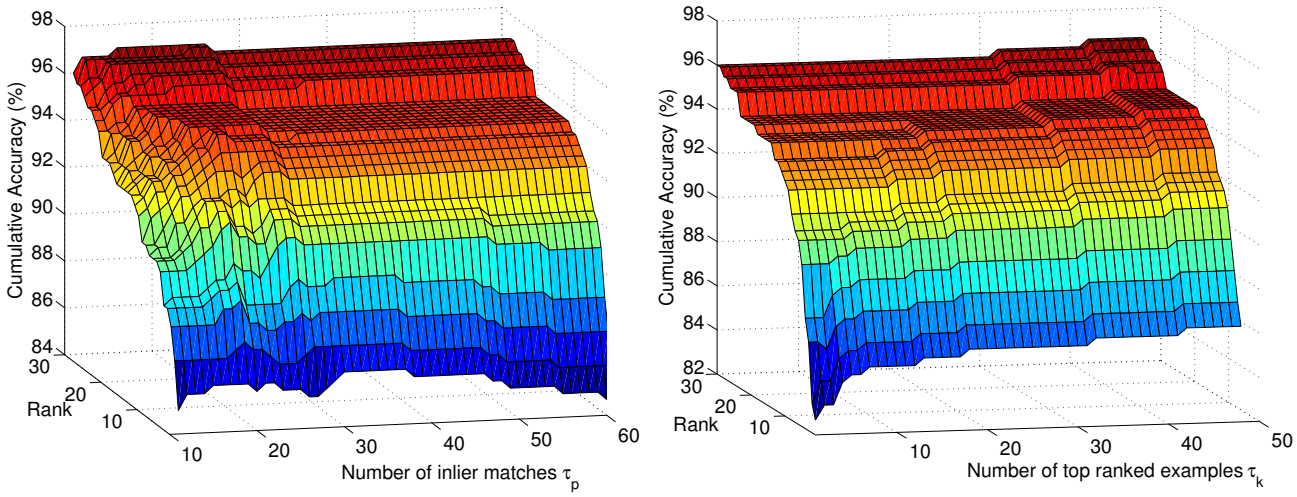


Figure 5: Comparison of the proposed semi-supervised ranking method using S^{3}VM algorithm with other appearance based methods on the TRECVID2003 dataset.

In addition, we also assess the performance of each component of the global features as well as the combined features. From the results shown in Figure 4, we can see that the approaches with the combined features clearly outperform the approaches with individual features. For the individual features, we found that the grid color moments method outperforms the other three methods.

5.3.2 *Performance of the S^{3}VM Method*

Finally, we compare the proposed semi-supervised ranking approach using the S^{3}VM method with other conventional appearance-based methods on global features, such as the approaches with color histogram [30] and color moments [31]. Note that we employ the Nearest Neighbor ranking results to select the most dissimilar examples as the negative samples for training S^{3}VM. Figure 5 shows the experi-



(a) different τ_p values (b) different τ_k values

Figure 6: Cumulative accuracy of NDK retrieval using NIM method on the TRECVID2003 dataset. (a). There is a wide range available from which to select the threshold value. The image pairs with below 30 inlier matches are viewed as non-duplicate in our experiments. (b) The overall accuracy grows with the number of top-K returns. We choose 50 as a trade-off between the accuracy and computational time.

mental results on the two datasets. Obviously, S^{3}VM significantly outperforms the color moment and color histogram methods. Specifically, S^{3}VM obtains about 33% improvement over the color moment method on the TRECVID2003 dataset. Compared with the supervised ranking methods including Nearest Neighbor ranking and SVM ranking, S^{3}VM achieves significantly better results, with around 10% improvement over the two conventional ranking methods.

5.4 Experiment II: Re-ranking with NIM on Local Features

5.4.1 Parameter Settings

The last key ranking stage for the MLR scheme is the NIM ranking using the proposed NDK matching technique. To deploy the NIM technique for the NDK retrieval task, we need to determine some parameter settings. In general, the total number of mesh vertices determines the computational complexity and the deformation accuracy of the NIM method. Empirically, we adopt a 14×16 mesh for all of our experiments. The regularization coefficient λ_r is set to 5×10^{-5} to allow large deformations. The order ν of the robust estimator is set to 4. The initial support is 100 and the decay rate is 0.5. We find the optimization of each NIM task requires around 9 iterations to achieve convergence.

5.4.2 Evaluation on the Choices of Two Thresholds

For the proposed NIM approach, there are two threshold parameters that can affect the resulting accuracy and efficiency performance. These are: (1) the minimal number of inlier matches for reporting positive NDKs, denoted by τ_p, and (2) the number of top ranked examples to be matched by NIM, denoted by τ_k.

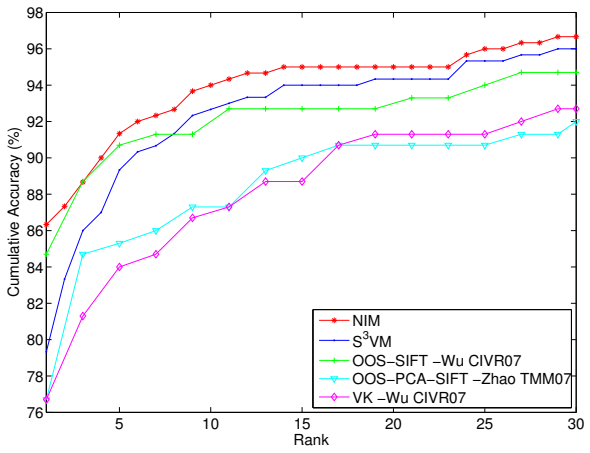
The first threshold parameter τ_p determines the threshold for predicting positive results. Normally, the smaller the value of τ_p, the higher the recall (the hit rate). At the same time, the precision is likely to drop with decreasing τ_p. Hence, it is important to determine an optimal threshold parameter. Although we do not have a theoretical approach to this, choosing a good τ_p value empirically seems not too difficult. To justify this, we evaluate the performance by varying the τ_p values. Figure 6(a) shows the surface of cumulative accuracies with the top 30 returned results on the TRECVID2003 dataset when τ_p varies from 10 to 50 (where τ_k is fixed to 50). From the results, we can see that good results can be obtained when setting the threshold τ_p between 15 and 30.

The second threshold parameter τ_k determines how many examples returned by the S^{3}VM ranking will be engaged for the NIM matching. Hence, it affects both the accuracy and efficiency performance. In general, the larger the value of τ_k, the more computational cost is incurred. However, τ_k value that is too small is likely to degrade the retrieval performance. Hence, choosing a proper τ_k value is important to balance the tradeoff between accuracy and efficiency performance. To see how τ_k affects the performance, Figure 6(b) shows the surface of cumulative accuracies with the top 30 returned results obtained by varying τ_k from 1 to 50 (with τ_p fixed to 30). From the results, we can see that the cumulative accuracy increases when τ_k increases and tends to converge when τ_k approaches 50. Therefore, in the rest of our experiments, we simply fix τ_k to 50 to achieve good efficiency. We will evaluate the efficiency performance in a subsequent part of this paper.

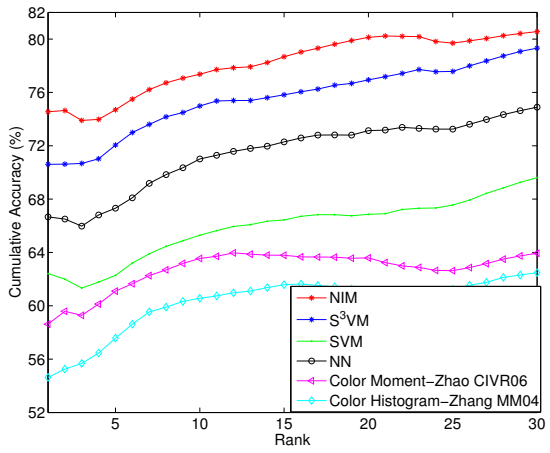
5.4.3 Comparisons of NDK Retrieval Performance

To examine the performance of the proposed NIM technique for retrieving NDKs, we compare our method with several state-of-the-art methods, including the OOS-SIFT method [27], the OOS-PCA-SIFT method [32], and the Visual Keywords (VK) methods [32]. Figure 7 shows the experimental results of the cumulative accuracy of the top 30 returned keyframes on both datasets.

For the TRECVID2003 dataset, it is relatively small and widely used as a benchmark testbed for NDK retrieval in literature. From the experimental results, we can draw several observations. First of all, the proposed S^{3}VM method with



(a) TRECVID2003 Dataset



(b) TRECVID2004 dataset

Figure 7: Comparison of cumulative accuracy of NDK retrieval results on the TRECVID2003 dataset (600 images) and the TRECVID2004 dataset (7006 images), respectively.

global features outperforms the OOS-PCA-SIFT method [32] and the VK method [27], which use local features. This again validates the effectiveness of the proposed semi-supervised ranking technique with S^3VM. Second, the proposed NIM algorithm with local features is significantly better than the S^3VM method. In particular, NIM achieves more than 8% improvement on the rank one accuracy over S^3VM. Finally, among all compared methods, the proposed NIM method achieves the best performance, outperforming the state-of-the-art OOS-SIFT method [27].

Turning next to the TRECVID2004 dataset, due to its large size, we have a difficulty of comparing our method with other existing methods, such as the OOS-SIFT and OOS-PCA-SIFT methods, which are computationally very intensive. Therefore, we only compare our method with some conventional approaches. Figure 7(b) shows the experimental results on the TRECVID2004 dataset. Similar to the previous dataset, NIM achieves the best performance among all the compared methods on this dataset. For other compared methods, S^3VM performs significantly better than both supervised SVM and NN methods.

5.5 Evaluation of Computational Cost

Finally, we empirically examine the efficiency performance of the proposed NIM and S^3VM methods. Both the global appearance features and local features are extracted offline. Table 1 and Table 2 summarize the overall computational time for comparing all pairs of keyframes on both datasets. From the results, we can see that NIM is more efficient than the OOS-SIFT method [27] and less efficient than the VK method which simply computes the similarity of visual words. Note that VK method often requires much preprocessing time cost for extracting the visual keywords offline. In addition, we clearly see that the methods using global features are significantly more efficient than the ones using local feature matching. This again validates the importance and effectiveness of the proposed multi-level ranking scheme for improving the efficiency. Finally, we also plot the computational cost and retrieval accuracy with respect to the number of top ranked examples (τ_k) to be compared by NIM in Figure 8. The results show that the larger the value of τ_k, the higher the computational cost and the better the matching accuracy. In particular, we found that the cumulative accuracy tends to converge to the best result when τ_k approaches to 50. In real-world applications, one can choose an appropriate τ_k to balance the tradeoff between accuracy and efficiency. For example, when τ_k equals to 10, each query for NIM takes about 1 second and achieves rather high cumulative accuracy, about 93%.

Table 1: Comparison of overall time cost of 300 queries on the TRECVID2003 dataset.

NIM	S^3VM	NN	OOS [27]	VK [27]
15.8min	3sec	1sec	6.5hour	1.5min

Table 2: Comparison of overall time cost of 1,953 queries on the TRECVID2004 dataset.

NIM	S^3VM	NN	OOS [27]	VK [27]
103.5min	8.1min	30sec	N/A	N/A

6. CONCLUSIONS

This paper proposed a novel nonrigid image matching method for Near-Duplicate Keyframe (NDK) retrieval. In contrast to traditional approaches with either projective geometry or bipartite graph matching, the proposed nonrigid image matching (NIM) algorithm recovers the explicit nonrigid mapping between two NDKs and effectively finds out the correct correspondences by a robust coarse-to-fine optimization scheme. Moreover, our method not only can detect the NDK pairs accurately, but also can recover the local deformations between them simultaneously. To further reduce the overall computational cost, we proposed an effective multi-level ranking scheme together with a semi-supervised ranking technique using semi-supervised SVM (S^3VM) to improve the ranking performance with the unlabeled data. We conducted extensive evaluations on two testbeds extracted from the TRECVID corpora. The promising experimental results showed that our method is importantly more effective than conventional approaches, especially for dealing with cases involving viewpoint changes and local deformations, which are very common in practice.

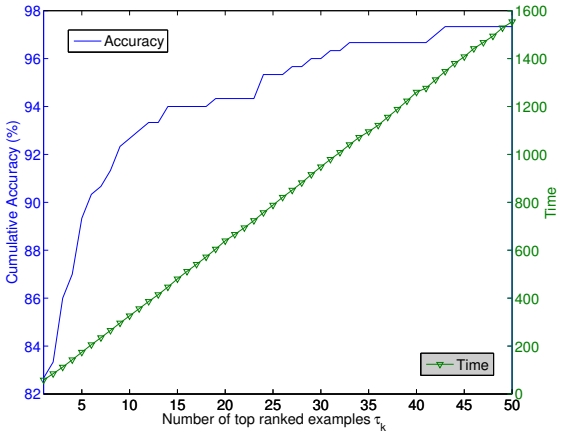


Figure 8: Computational efficiency and retrieval performance on the TRECVID2003 dataset. The *left* vertical axis shows mean *cumulative accuracy* of the top 30 returned results, and the *right* vertical axis represents the overall time cost for all 300 queries.

Acknowledgments

The authors would like to thank Prof. C.W. Ngo and Mr. W.L. Zhao for providing their results and the fruitful discussions. The work was fully supported by the Research Grants Council Earmarked Grant (CUHK4150/07E) and the Singapore MOE AcRF Tier-1 research grant (RG67/07).

7. REFERENCES

[1] `http://vireo.cs.cityu.edu.hk/research/NDK/ndk.html`.
[2] `http://www.cse.cuhk.edu.hk/~jkzhu/dup_detect.html`.
[3] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *Proc. European Conf. Computer Vision*, pages 404–417, 2006.
[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
[5] J. Canny. A computational approach to edge detection. *IEEE Trans. PAMI*, 8(6):679–698, 1986.
[6] O. Chum and J. Matas. Matching with prosac- progressive sample consensus. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 220–226, 2005.
[7] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, 1981.
[8] P. Fua and Y. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int'l J. Computer Vision*, 16(1):35–56, Sep. 1995.
[9] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Professional, Inc., 1990.
[10] C.-H. Hoi, W. Wang, and M. R. Lyu. A novel scheme for video similarity detection. In *CIVR*, pages 373–382, 2003.
[11] S. C. Hoi and M. R. Lyu. A multi-modal and multi-level ranking framework for content-based video retrieval. *To appear in IEEE Transactions on Multimedia*, 2008.
[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int'l J. Computer Vision*, 1(4):321–331, Jan. 1988.
[13] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detction and sub-image retrieval system. In *ACM MULTIMEDIA'04*, pages 869–876. ACM, 2004.
[14] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42(5):300–311, 1993.
[15] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int'l J. Computer Vision*, 60(2):91–110, 2004.
[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
[17] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang. Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In *ACM MULTIMEDIA'06*, pages 845–854. ACM, 2006.
[18] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. 29(1):51–59, January 1996.
[19] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration, and realistic augmentation. *Int'l J. Computer Vision*, 76(2):109–122, 2008.
[20] A. Qamra, Y. Meng, and E. Y. Chang. Enhanced perceptual distance functions and indexing for image replica recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):379–391, 2005.
[21] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML'05*, pages 824–831. ACM Press, 2005.
[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision (ICCV2003)*, pages 1470–1477, 2003.
[23] TRECVID. TREC video retrieval evaluation. In *http://www-nlpir.nist.gov/projects/trecvid/*.
[24] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
[25] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *ACM MULTIMEDIA'07*, pages 168–177. ACM, 2007.
[26] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *ACM MULTIMEDIA'07*, pages 218–227. ACM, 2007.
[27] X. Wu, W.-L. Zhao, and C.-W. Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *ACM CIVR'07*, pages 162–169. ACM, 2007.
[28] Z. Xu, R. Jin, J. Zhu, I. King, and M. R. Lyu. Efficient convex relaxation for transductive support vector machine. In *NIPS'2007*, 2007.
[29] R. Yan, A. G. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MULTIMEDIA'03*, pages 343–346, 2003.
[30] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM MULTIMEDIA'04*, pages 877–884. ACM, 2004.
[31] W. Zhao, Y. Jiang, and C. Ngo. Keyframe retrieval by keypoints: Can point-to-point matching help? In *CIVR06*, pages 72–81, 2006.
[32] 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. on Multimedia*, 9(5):1037–1048, 2007.
[33] J. Zhu. Semi-supervised learning literature survey. Technical report, Carnegie Mellon University, 2005.
[34] J. Zhu, S. C. Hoi, and M. R. Lyu. Face annotation by transductive kernel fisher discriminant. *IEEE Trans. on Multimedia*, 10(1):86–96, 2008.
[35] J. Zhu and M. R. Lyu. Progressive finite newton approach to real-time nonrigid surface detection. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2007.
[36] J. Zhu, M. R. Lyu, and T. S. Huang. A fast 2d shape recovery approach by fusing features and appearance. *To appear in IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.