Near-Duplicate Keyframe Retrieval by Semi-Supervised Learning and Nonrigid Image Matching

JIANKE ZHU, Zhejiang University
STEVEN C. H. HOI, Nanyang Technological University
MICHAEL R. LYU, The Chinese University of Hong Kong
SHUICHENG YAN, National University of Singapore

4

Near-duplicate keyframe (NDK) retrieval techniques are critical to many real-world multimedia applications. Over the last few years, we have witnessed a surge of attention on studying near-duplicate image/keyframe retrieval in the multimedia community. To facilitate an effective approach to NDK retrieval on large-scale data, we suggest an effective Multi-Level Ranking (MLR) scheme that effectively retrieves NDKs in a coarse-to-fine manner. One key stage of the MLR ranking scheme is how to learn an effective ranking function with extremely small training examples in a near-duplicate detection task. To attack this challenge, we employ a semi-supervised learning method, semi-supervised support vector machines, which is able to significantly improve the retrieval performance by exploiting unlabeled data. Another key stage of the MLR scheme is to perform a fine matching among a subset of keyframe candidates retrieved from the previous coarse ranking stage. In contrast to previous approaches based on either simple heuristics or rigid matching models, we propose a novel Nonrigid Image Matching (NIM) approach to tackle near-duplicate keyframe retrieval from real-world video corpora in order to conduct an effective fine matching. Compared with the conventional methods, the proposed NIM approach can recover explicit mapping between two near-duplicate images with a few deformation parameters and find out the correct correspondences from noisy data simultaneously. To evaluate the effectiveness of our proposed approach, we performed extensive experiments on two benchmark testbeds extracted from the TRECVID2003 and TRECVID2004 corpora. The promising results indicate that our proposed method is more effective than other state-of-the-art approaches for near-duplicate keyframe retrieval.

Categories and Subject Descriptors: I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms: Algorithms, Performance, Experimentations

Additional Key Words and Phrases: Near-duplicate keyframe, image copy detection, nonrigid image matching, semi-supervised learning

ACM Reference Format:

Zhu, J., Hoi, S. C. H., Lyu, M. R., and Yan, S. 2011. Near-duplicate keyframe retrieval by semi-supervised learning and nonrigid image matching. ACM Trans. Multimedia Comput. Commun. Appl. 7, 1, Article 4 (January 2011), 24 pages.
DOI = 10.1145/1870121.1870125 http://doi.acm.org/10.1145/1870121.1870125

A short version of this article was accepted by ACM Multimedia 2008.
The work was supported in part by three grants: the Research Grants Council General Research Fund (CUHK4154/09E), the Singapore MOE AcRF Tier-1 research grant (RG67/07), and NRF/IDM Program under research Grant NRF2008IDMIDM004-029.
Authors' addresses: J. Zhu, Zhejiang University; email: jianke.zhu@gmail.com; S. C. H. Hoi, Block N4, Room 02c-112, Division of Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore 639798; M. R. Lyu, Department of Computer Science and Engineering, Chinese University of Hong Kong, China; S. Yan, ECE Department, National University of Singapore, Singapore 117576.

DOI 10.1145/1870121.1870125 http://doi.acm.org/10.1145/1870121.1870125

1. INTRODUCTION

Near-Duplicate Keyframes (NDK) detection and retrieval techniques are beneficial for many real-world applications, such as copyright infringement detection [Ke et al. 2004; Qamra et al. 2005], elimination of near-duplicates from Web video search results [Wu et al. 2007b], and news video search [Smeaton et al. 2006]. In general, NDK refers to a pair of keyframes in a video corpus, for which the two keyframes of the pair are closely similar to each other apart from minor differences mostly due to the variations of rendering conditions, capturing conditions, or editing operations [Zhang and Chang 2004; Zhao et al. 2007]. NDKs are very common for video retrieval tasks, especially for broadcast videos [Wu et al. 2007b; Zhang and Chang 2004; Zhao et al. 2007]. Besides the reason of duplicate or near-duplicate videos, another reason is imperfect video segmentation results. For example, a naive uniform segmentation approach often leads to generating mass NDKs in the segmentation results. Even for the state-of-the-art automatic segmentation techniques, NDK is still commonly found in the segmentation results due to the difficulties of setting very accurate parameters in practice.

Due to some well-known factors, NDK retrieval is a challenging research problem. One is that videos may be captured by different devices with quite different hardware under a variety of illumination conditions. Additionally, video editing often produces extra geometric and photometric transformations and occludes the original video by adding captions. Furthermore, there is still a lack of an effective feature extraction method to represent NDK, while lots of feature descriptors have already been developed for object recognition [Mikolajczyk and Schmid 2005; Everingham et al. 2007] and face recognition [Zhao et al. 2003]. Figure 1 shows some example pairs of duplicate keyframes extracted from the TRECVID2003 video corpus [Smeaton et al. 2006].

During the past several years, there has been a surge of research attention on copy-detection and NDK retrieval in the multimedia community [Ke et al. 2004; Qamra et al. 2005; Wu et al. 2007b; Wu et al. 2007c; Zhang and Chang 2004; Zhao et al. 2007; Chum et al. 2007; Xu et al. 2008; Zhu et al. 2008]. Some methods directly extend the conventional content-based image retrieval (CBIR) techniques for the NDK detection and retrieval task; these approaches often employ the statistical information extracted from the whole image, that is, color histogram and color moment [Qamra et al. 2005; Zhang and Chang 2004]. Although usually deemed very efficient in finding identical copies, these methods may neglect the spatial information, and become not very effective for real NDKs as they often fail to address the variations of viewpoint changes, illumination changes, partial occlusions and image editing.

Instead of extracting the features from the whole image, some alternative approaches using local feature point matches can deal with some geometric transformations and illumination variations by taking advantage of the recent advances in local feature descriptors [Mikolajczyk and Schmid 2005]. The major drawback of these approaches is that they often incur heavy computational cost in finding feature correspondences. Nevertheless, some efficient solutions have been proposed. For example, Ke et al. [2004] proposed an efficient method using locality-sensitive hashing indexing for PCA-SIFT features. However, their method often assumes a *rigid* projective geometry transformation between NDKs, which may suffer from some outlier matches due to object movements and focal length changes. To relax the strong rigid projection assumption, Zhang and Chang [2004] proposed a stochastic Attributed Relational Graph (ARG) matching framework, which involves a computationally intensive process of stochastic belief propagation. In most recent work, Zhao et al. [2007] proposed a one-to-one symmetric (OOS) matching method, in which a local smoothing constraint is applied to remove the outlier matches. In Ngo et al. [2006], Pattern Entropy (*PE*) rather than the total number of inlier matches is employed as a similarity measure for OOS method, which achieves the state-of-art performance on TRECVID2003 dataset. Similar to other *bipartite graph matching* methods, the OOS method considers

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

Fig. 1. Some near-duplicate keyframes examples from TRECVID2003 video corpus. 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).

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, partial occlusions and image zooming lead to large *PE*, in which $PE \leq 0.5$ is considered as an NDK pair [Ngo et al. 2006].

To facilitate an effective approach to NDK retrieval on large-scale data, we try to attack this challenge under a Multi-Level Ranking (MLR) framework, which integrates three different ranking components into a unified solution: nearest neighbor ranking, semi-supervised ranking, and NIM-based ranking. As the image feature representation is essential to nearest neighbor ranking and machine learning-based ranking methods, we also explore the effective feature extraction methods. In this article, five kinds of features are investigated: GIST [Oliva and Torralba 2001], grid color moment, Gabor wavelets transform [Lades et al. 1993], Local Binary Pattern [Ojala et al. 1996], and edge histogram.

Considering that the previous approaches employ either rigid projective models [Ke et al. 2004] or bipartite graph matching, we propose a novel *Nonrigid Image Matching* (NIM) method for near-duplicate keyframe retrieval in this paper. Unlike the previous conventional approaches, we assume that there may exist nonrigid transformations between the two NDKs. The key to tackle the NIM problem is to utilize an iterative coarse-to-fine optimization scheme to progressively reject the outliers, which takes advantage of a closed-form solution for a given set of local feature matches. As our method takes into account local deformations, it often obtains more inlier correspondences than conventional rigid projective geometry models and the OOS graph matching method; this characteristic plays a very important 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 [Lowe 2004].

Compared with the previous approaches, our 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 NDKs, which is often not obtainable by conventional methods. This attractive feature is important for part-based or subimage detection and retrieval. Additionally, our method is rather efficient, with a processing speed of about ten pairs of keyframes per second in a regular PC with moderate configuration.

In summary, this paper includes four 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. Second, to enable the proposed technique to be applicable to large-scale applications, we suggest a *Multilevel 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 [Hoi et al. 2003; Hoi and Lyu 2008], 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) to improve 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. Finally, we propose a very effective feature representation scheme for NDK, which is one of the keys to achieve the excellent performance.

The rest of this article is organized as follows. Section 2 reviews some existing approaches for NDK retrieval. Section 3 presents a multilevel ranking scheme together with a semi-supervised SVM method for NDK retrieval. Section 4 proposes the nonrigid image matching method for detecting NDK with local feature correspondences. 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 [Ke et al. 2004; Qamra et al. 2005; Wu et al. 2007a; Wu et al. 2007c; Zhang and Chang 2004]. Generally, most of these existing approaches can be roughly categorized into two groups: *appearance-based methods* and *local feature-based methods*.

The appearance-based methods often measure the similarity between two keyframes using the extracted global visual features, such as color histogram [Zhang and Chang 2004] and color moments [Zhao et al. 2006]. As the feature is formed by the global statistical information, the associated spatial information is lost. A remedy to alleviate this issue is to introduce the grid representation [Zhao et al. 2006; Zhu et al. 2008b]. As keyframes are often compactly represented in the vector space, these appearance-based methods are advantageous for their high efficiency, and thus can take advantage of conventional CBIR methods and mature data indexing techniques [Qamra et al. 2005]. The major drawback for these methods is that they are often not very robust to partial occlusions, illumination changes, and some geometric transformations.

On the other hand, the local feature-based methods detect local salient points in two keyframes and measure their similarity by counting the number of inlier matches between two keypoint sets. Usually, keypoints are the salient regions detected over image scales and their descriptors are often invariant to certain variations and transformations. They overcome the limitations of the global appearance-based methods, and thus often achieve better performance [Ke et al. 2004; Zhao et al. 2007]. But they may incur a heavy computational cost for the matching of two keypoint sets, which may contain more than one thousand of keypoints. To reduce the total number of local feature matching between NDKs, nearest neighbor filtering is performed to shrink the size of candidate list. On the other hand, the bag-of-words method [Wu et al. 2007c] is also introduced to NDK retrieval task, which wins success in object recognition [Everingham et al. 2007].

Recently, local feature-based methods have been actively studied. Ke et al. [2004] 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. Sivic and Zisserman [2003] proposed the local keypoints approach for object matching and retrieval in movies. Zhao et al. [2007] 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.

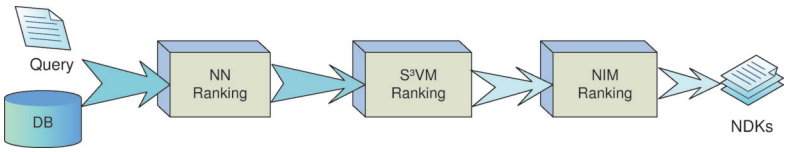


Fig. 2. A multilevel ranking framework for NDK retrieval.

Most of these methods fall in the same category of point-to-point bipartite graph matching. In the most recent work, Xu et al. [2008] proposed a Spatially Aligned Pyramid Matching approach, which tends to robustly handle spatial shifts and scale changes. The first matching stage calculates the pairwise distances between blocks from the input image using SIFT features and Earth Movers Distance [Rubner et al. 2000], and the second stage handles the piecewise spatial shifts and scale variation using a modified Earth Movers Distance.

In this article, the proposed NIM technique goes beyond the conventional point-to-point bipartite graph matching methods. In contrast to the conventional techniques, our method not only can recover the explicit mapping between two near-duplicate keyframes with nonrigid transformation models, but also can effectively find the correct correspondences from noisy data. Though similar techniques are actively studied for tracking in computer vision and graphics [Zhu and Lyu 2007; Zhu et al. 2008c; Pilet et al. 2008; Zhu et al. 2009], to the best of our knowledge, we are the first to study nonrigid image matching comprehensively for NDK retrieval tasks.

3. MULTILEVEL FRAMEWORK OF NEAR-DUPLICATE KEYFRAME RETRIEVAL

3.1 Framework Overview

Since directly applying local feature matching based method to large-scale applications could be computationally intensive, we employ a Multi-Level Ranking (MLR) framework for efficiently tackling the NDK retrieval task. This framework is able to greatly improve the efficiency and scalability of our proposed solution. This strategy has been widely used, which is also shown to be successful in multimedia retrieval [Hoi et al. 2003; Hoi and Lyu 2008]. As shown in Figure 2, the proposed MLR framework attacks the NDK retrieval task in a coarse-to-fine ranking manner. Specifically, our multilevel ranking scheme integrates three different ranking components.

—*Nearest Neighbor Ranking (NNR).* This is to rank the keyframes with simple nearest neighbor search on the global features.

—*Semi-Supervised Ranking (SSR).* This is to rank the keyframes with a semi-supervised ranking method. Note that we employ the pseudo relevance feedback technique, in which SSR chooses a short list of the most dissimilar examples from NNR results as the negative examples.

—*Nonrigid Image Matching (NIM).* This is to rank the keyframes by applying the proposed NIM method. We re-rank the top retrieved results from SSR using NIM, where SSR greatly reduces the searching spaces for NIM.

In summary, the first two ranking components are based on global features for efficiently filtering out the irrelevant results, and the last component provides a fine reranking based on the local features. This scheme makes the proposed NIM solution applicable to large-scale real applications.

3.2 Formulation: From a Machine Learning Viewpoint

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 [Yan et al. 2003]. Specifically, we can conduct a query-by-example retrieval for ranking the unlabeled data in $\mathcal{G}$ according to their distances from the examples in the query set. Then we select a short list of the 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 [Vapnik 1998], 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:

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

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

While SVM can be directly applied to solve the learning task, its performance may be poor when there are only a limited number of labeled examples. In fact, this is the case of an NDK retrieval task, as extremely few positive examples will be provided for each query. To overcome the second difficulty, we next introduce a semi-supervised learning technique for exploring both labeled and unlabeled data for the retrieval tasks.

3.3 Semisupervised 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 [Xu et al. 2007; Zhu 2005; Zhu et al. 2008a]. In this article, 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 [Sindhwani et al. 2005].

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 [Sindhwani et al. 2005]. Let $\mathcal{H}$ denote the original Hilbert space reproduced by kernel function $k(\cdot, \cdot)$, and $\widetilde{\mathcal{H}}$ denote the deformed Hilbert space. Sindhwani et al. [2005] assume the following relationship between the two Hilbert spaces:

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

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 follows:

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

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 $L = \text{diag}(S\mathbf{1}) - S$, where $\mathbf{1}$ denotes a vector with all one elements. Moreover, $S \in R^{n\times n}$ is a similarity matrix and each element $S_{i,j}$ is calculated by:

$$S_{ij} = S_{ji} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\varsigma^2}}, & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are adjacent,} \\ 0, & \text{otherwise,} \end{cases}$$

where ς denotes the kernel width for a graph Laplacian. Various similarity measures can be used to build the adjacent matrix, such as L_1 norm, L_2 norm, and cosine similarity.

Consequently, the new kernel k can be formulated as follows:

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) - \kappa_{\mathbf{y}}^\top (I + MK)^{-1} M\kappa_{\mathbf{x}}. \tag{2}$$

Replacing the kernel k in Equation (1) by the kernel $\tilde{k}$ in Equation (2), we can train the semi-supervised SVM classifier. That is to say, we employ the deformed kernel to train a regular SVM, in which the query NDK and the selected pseudo negative examples are treated as the labeled data. Note that Equation (2) can also be used to compute the kernel for transductive learning, and the new deformed kernel matrix $\tilde{K} \in R^{n\times n}$ can be derived as follows:

$$\tilde{K} = K - \mathcal{K}(I + MK)^{-1}MK. \tag{3}$$

It can further be simplified through Kailath Variant:

$$\tilde{K} = (I + KM)^{-1}K.$$

In addition, we have the further equality

$$\tilde{K} = K(I + MK)^{-1}. \tag{4}$$

From the previous equations, we summarize the complete S^{3}VM reranking algorithm into Figure 3.

4. 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.

4.1 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

Algorithm 1 Semi-Supervised SVM Re-ranking

Input

—X: the extracted features of all images in the dataset
—k: an input kernel function
—$\mathbf{x}_q$: a query example (may contain multiple images)
—n_-: number of pseudo-negative examples to be used

Procedure

/* offline computation before handling a query */
1: Calculate initial kernel matrix K: $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$
2: Compute graph Laplacian L and semi-definite positive matrix M
3: Calculate the semi-supervised kernel matrix $\tilde{K}$ by:

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

/* online computation for handling a query $\mathbf{x}_q$ */
4: for i = 1 to n
5: $dist(\mathbf{x}_i) = \mathsf{euclidean}(\mathbf{x}_q, \mathbf{x}_i)$; /* euclidean distance */
6: end for
7: $\mathcal{L}^- = \mathsf{top\_k\_max\_dist}(\text{dist}, n_-)$; /* get top n_- most dissimilar examples */
8: $\mathcal{L} = \mathcal{L}^- \cup \{\mathbf{x}_q\}$
9: $\boldsymbol{\alpha} = \mathsf{SVM\_solver}(\mathcal{L}, \tilde{K})$ /* train an SVM classifier with the semi-supervised kernel */
10: for i = 1 to n
11: $f_{\mathrm{SVM}}(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathcal{L}} \alpha_j y_j \tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$; /* SVM distance */
12: end for
13: $\mathcal{R} = \mathsf{top\_k\_max\_dist}(f_{\mathrm{SVM}}, m)$; /* rank top m most relevant examples by SVM*/

Output

—$\mathcal{R}$: the rank list output by the SVM ranking.

End

Fig. 3. Semi supervised SVM reranking algorithm.

data simultaneously. This method has been successfully applied to real-time nonrigid surface tracking in computer vision [Pilet et al. 2008; Zhu and Lyu 2007; Zhu et al. 2009]. Unlike the traditional nonrigid image registration, the NIM method is a fully automatic solution and does not require manual initialization.

The main idea of NIM is to recover the local deformations from the salient feature matches between the two images and to reject the outlier matches simultaneously. Thus, we can simply choose the total number of inlier matches τ as a confidence measure to decide whether the two keyframes are near-duplicate or not. More specifically, given a set of correspondences $\mathcal{M}$ between the model and the input image built through a local feature matching algorithm, we manage to find the nonrigid mapping from these correspondences. 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

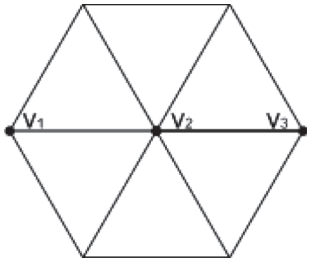


Fig. 4. Illustration of the regularization for an hexagonal unit in the mesh model.

vertices in the mesh. Thus, the mapping function $T_{\mathbf{s}}(\mathbf{m})$ is defined as follows:

$$T_{\mathbf{s}}(\mathbf{m}) = \begin{bmatrix} u_i & u_j & u_k \\ v_i & v_j & v_k \end{bmatrix} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \end{bmatrix}^{\top}, \tag{5}$$

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:

$$E_c(\mathbf{s}) = \sum_{\mathbf{m} \in \mathcal{M}} \omega_{\mathbf{m}} \mathcal{V}(\delta, \sigma), \tag{6}$$

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 follows:

$$\delta(\mathbf{m}) = \|\mathbf{m}_1 - T_{\mathbf{s}}(\mathbf{m}_0)\|. \tag{7}$$

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 [Boyd and Vandenberghe 2004]:

$$\mathcal{V}(\delta, \sigma) = \begin{cases} \frac{\delta(\mathbf{m})}{\sigma^{\nu}}, & \mathcal{M}_1 = \left\{\mathbf{m} | \delta(\mathbf{m}) \le \sigma^2\right\} \\ \sigma^{2-\nu}, & \mathcal{M}_2 = \overline{\mathcal{M}_1} \end{cases}, \tag{8}$$

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 larger the value of the support σ is, the larger the number of the correspondences is included. Conversely, when 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. In particular, we consider the following regularized objective function for energy minimization:

$$E(\mathbf{s}) = E_c(\mathbf{s}) + \lambda_r E_r(\mathbf{s}), \tag{9}$$

where $E_r(\mathbf{s})$ is the regularization term that represents the deformation energy, and λ_r is a regularization coefficient. Similar energy functions have been used in deformable surface fitting [Kass et al. 1988; Pilet et al. 2008; Zhu et al. 2008b].

The regularization term E_r in the given formulation, also known as "internal force" in Snakes [Kass et al. 1988], is composed of the sum of the squared second-order derivatives of the mesh vertex coordinates. Specifically, since the mesh is regular, the length differences between two colinear connected

edges is penalized, as illustrated in Figure 4. Thus, the regularization energy e_r [Fua and Leclerc 1995; Pilet et al. 2008] for each edge pair in the conlinear connected edges set $\mathcal{E}$ can be written as follows:

$$\begin{aligned} e_r &= [(u_1 - u_2) - (u_2 - u_3)]^2 + [(v_1 - v_2) - (v_2 - v_3)]^2 \\ &= ([\ 1\ \ -2\ \ 1\ ]^\top [\ u_1\ \ u_2\ \ u_3\ ])^2 + ([\ 1\ \ -2\ \ 1\ ]^\top [\ v_1\ \ v_2\ \ v_3\ ])^2. \end{aligned} \tag{10}$$

Let $\mathbf{h} \in R^N$ be an auxiliary vector containing a set of coefficients so that we can rewrite e_r with respect to the vertex coordinates:

$$e_r = (\mathbf{h}^\top \mathbf{u})^2 + (\mathbf{h}^\top \mathbf{v})^2 = \mathbf{u}^\top (\mathbf{h}\mathbf{h}^\top)\mathbf{u} + \mathbf{v}^\top (\mathbf{h}\mathbf{h}^\top)\mathbf{v}.$$

Furthermore, $E_r(\mathbf{s})$ can be formulated as the summation of all edge pairs:

$$E_r(\mathbf{s}) = \sum_{\mathcal{E}} \mathbf{u}^\top (\mathbf{h}\mathbf{h}^\top)\mathbf{u} + \mathbf{v}^\top (\mathbf{h}\mathbf{h}^\top)\mathbf{v} = \mathbf{s}^\top \begin{bmatrix} \mathcal{K} & 0 \\ 0 & \mathcal{K} \end{bmatrix} \mathbf{s}, \tag{11}$$

where $\mathcal{K} = \sum_{\mathcal{E}} \mathbf{h}\mathbf{h}^\top$ is a sparse and banded matrix, which is determined by the structure of the explicit mesh model.

4.2 Optimization

We now present an effective technique to solve the optimization task as shown in Equation (9). Since the robust estimator function in Equation (8) is nonconvex, it 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 [Zhu and Lyu 2007; Zhu et al. 2009]. Given a set of inlier matches $\mathcal{M}_1$, the solution for the optimization problem in Equation (9) can be obtained through solving the following two linear equations via LU decomposition:

$$(\lambda_r \mathcal{K} + A)\mathbf{u} = \mathbf{b}_u \tag{12}$$

$$(\lambda_r \mathcal{K} + A)\mathbf{v} = \mathbf{b}_v, \tag{13}$$

where $A \in R^{N \times N}$ is computed as follows:

$$A = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^v} \mathbf{t}\mathbf{t}^\top \tag{14}$$

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

$$\mathbf{b}_u = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^v} u\mathbf{t} \quad \text{and} \quad \mathbf{b}_v = \sum_{\mathbf{m} \in \mathcal{M}_1} \frac{\omega_{\mathbf{m}}}{\sigma^v} v\mathbf{t}, \tag{15}$$

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

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

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 given 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

Algorithm 2 Algorithm of Nonrigid Image Matching (NIM)
Input

—Parameters: ν, η, λ_r, σ_0
—Query image

Pre-compute
1: Build mesh model $\mathbf{s}_0$ for a query image
2: Compute $\mathcal{K}$ and barycentric coordinates (ξ_1, ξ_2, ξ_3) for each keypoint $\mathbf{m}_0$
Nonrigid Image matching
For an image in the gallery set:
Select the active set by modified RANSAC
Set $\sigma = \sigma_0$
Repeat

(1) Compute A and $\mathbf{b}$ by Eqn. 14 and Eqn. 15
(2) Solve the linear system: Eqn. 12 and Eqn. 13
(3) Calculate residual error δ by Eqn. 7 and the inlier set $\mathcal{M}_1$ by Eqn. 8
(4) Update $\sigma = \eta \cdot \sigma$

Until converge!
Output

—the total number of inlier matches $\tau = |\mathcal{M}_1|$
—the mesh vertices $\mathbf{s}$

End

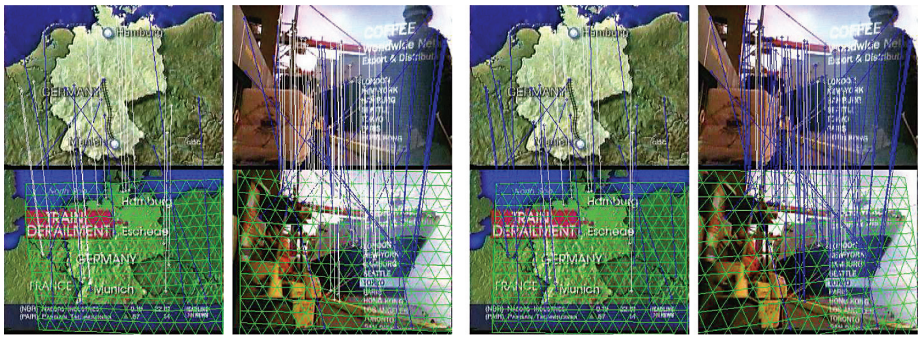
Fig. 5. The Nonrigid Image Matching (NIM) algorithm.

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 [Chum and Matas 2005; Fischler and Bolles 1981] 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 usually requires fewer steps empirically. From all of these, we summarize the details of the nonrigid image matching (NIM) algorithm in Figure 5.

4.3 Local Feature Matching

Interest point detection and matching is a fundamental research problem in computer vision. Many effective approaches have been proposed in literature. One of the most widely used methods is SIFT [Lowe 2004], which computes a histogram of local oriented gradients around the interest point and stores the bins in a 128-dimensional vector. To improve SIFT, Ke et al. [2004] proposed an extended method by applying Principle Component Analysis [Fukunaga 1990] on the gradient image, which



(a) Nonrigid image matching (b) Rigid matching

Fig. 6. Examples of our detection results on view point changes. (a) We overlaid the estimated mesh onto the test image. (b) The model mesh is mapped to the input image using homography estimated by rigid matching. The white lines are the inlier matches, and the blue lines (black lines on the grayscale printed paper) are outlier matches. We use the same notation in the following examples.

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 [Mikolajczyk and Schmid 2005], and is also slower than the original SIFT in the feature extraction. Instead of using SIFT or PCA-SIFT, we adopt SURF [Bay et al. 2006], 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 SIFT, it is more compact and hence reduces the computational cost for keypoint matching.

4.4 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. To compare the conventional methods based on the rigid projective geometry assumption, we also conduct the experiments with the same set of feature correspondences using RANSAC fitting. Specifically, we employ the homography estimation technique [Chum et al. 2007], which is typically used in the pairwise image matching. In the experiment, we also take advantage of the feature matching score to reduce the total number of RANSAC trails, and set the inlier threshold to two pixels. To make comparison with the NIM results, we map the model mesh in the reference image onto the input image with estimated homography matrix.

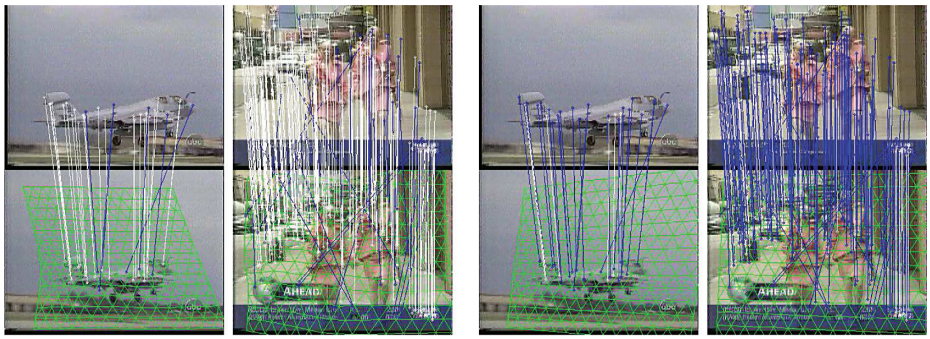
Figures 6–11 show some examples of the detection results using NIM and rigid matching for the various kinds of NDKs. All results on the duplicate pairs from Columbia's TRECVID2003 dataset can be found at the Web page.[1] In particular, the proposed NIM technique can effectively detect a variety of NDKs 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.

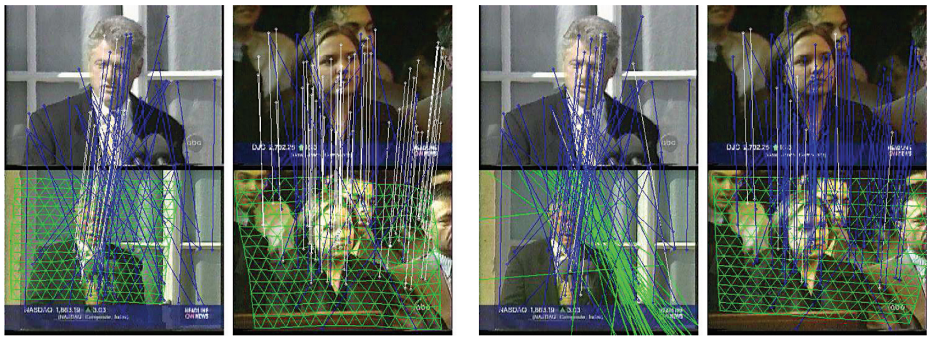
—*Lens change.* This case is caused by the changes of camera lens, such as zooming in or zooming out.

[1]http://www.cse.cuhk.edu.hk/~jkzhu/dup_detect.html.



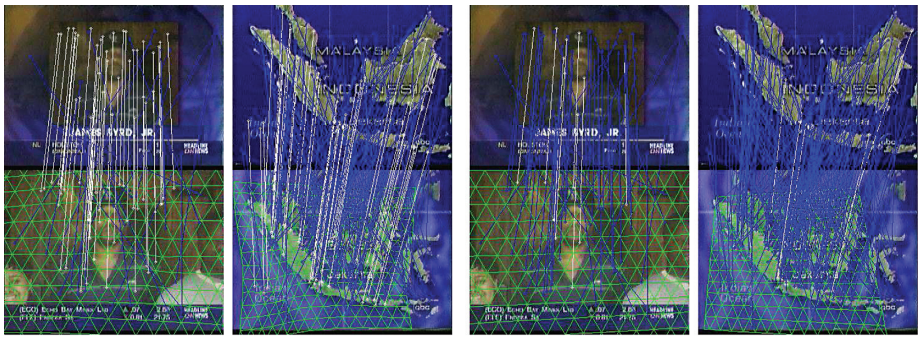
(a) Nonrigid image matching (b) Rigid matching

Fig. 7. Examples of our detection results on the keyframes with object movements.



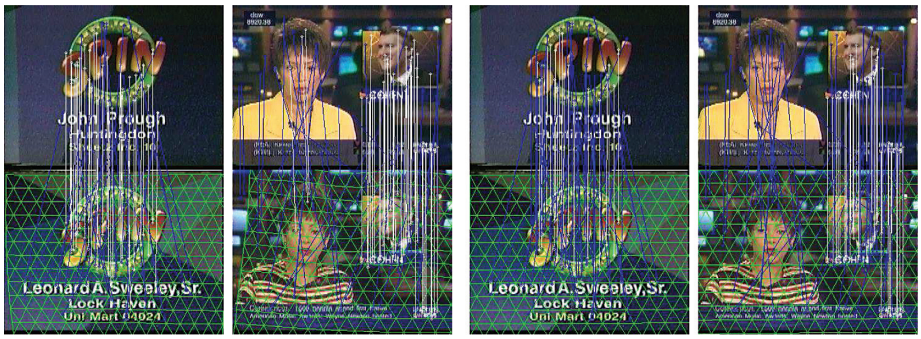
(a) Nonrigid image matching (b) Rigid matching

Fig. 8. Examples of our detection results on the lens changes.



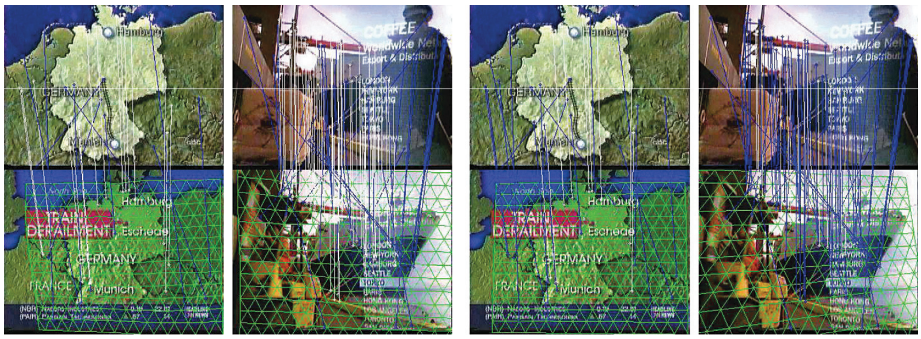
(a) Nonrigid image matching (b) Rigid matching

Fig. 9. Examples of our detection results on the subimage duplicates.



(a) Nonrigid image matching (b) Rigid matching

Fig. 10. Examples of our detection results on the keyframes with small regional changes.



(a) Nonrigid image matching (b) Rigid matching

Fig. 11. Examples of our detection results on the keyframes with partial occlusions.

—*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.

From these results, we can clearly see that NIM obtains more inliers than rigid matching. This is because that most of the NDK pairs contain the spatial deformations due to object movements or different capture conditions. These deformations can be captured by the nonrigid mapping in the most cases. Being the local feature-based methods, both NIM and rigid image matching are effective for partial occlusions and the subimages.

5. EXPERIMENTS

In this section, we perform our empirical evaluation on the proposed techniques for NDK retrieval task. Two key techniques in our proposed approach are studied comprehensively in the experiments. In the first experiment, we examine the effectiveness of the Multilevel Ranking scheme for filtering out the irrelevant results. In particular, we would like to study 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 [Zhang and Chang 2004], which consists of 600 keyframes with 150 near duplicate image pairs and 300 nonduplicate images extracted from the TRECVID2003 corpus [Zhang and Chang 2004]. All the keyframes are with the same size, 352×264. The other is CityU's TRECVID2004 dataset[2] recently collected by Ngo et al. [2006]. 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.

In our experiments, we adopt the evaluation protocol used in Zhao et al. [2007] in order to make a fair comparison with the state-of-the-art approaches. More specifically, all NDK pairs are treated 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 [Zhao et al. 2007], 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. More specifically, *cumulative accuracy* is defined as the ratio between the number of correctly retrieved NDKs and the ground truth of total number of NDKs in the top k returned the keyframes. As a yardstick for assessing the performance, we compare our method with the recently proposed OOS matching algorithm [Zhao et al. 2007], 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 $\gamma_I = 10^{-1}$. Moreover, the Laplacian graph is constructed based on heat kernel.

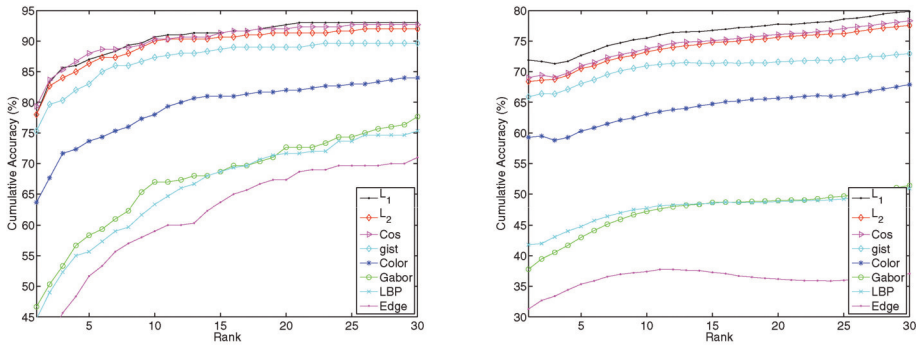
All the experiments in this paper were carried out on a notebook computer with Intel Core-2 Duo 2.0GHz processor and 2GB RAM. All the proposed methods are implemented in Matlab, for which some routines are written in C code. The code can be downloaded from the Web page.[3]

5.2 Image Representation

To facilitate the effective Nearest Neighbor Ranking and Semi-Supervised Ranking, we need represent the images by the feature vectors, which is a key step for NDK retrieval. In the past decade, the global feature representation techniques have been extensively studied in image processing and CBIR community. Also, a wide variety of global feature extraction techniques were proposed. Comparing to the four kinds of features used in Zhu et al. [2008], we also add the GIST feature in this paper. Therefore, we extract following five kinds of effective global features:

5.2.1 *GIST Feature.* GIST feature [Oliva and Torralba 2001] is an effective tool for representing scene structure by the spatial envelope, and obtains state-of-art performance in scene recognition. GIST is based on a set of 2D Fourier Transform of the input image, which is partitioned into 4×4 grid.

[2]http://vireo.cs.cityu.edu.hk/research/NDK/ndk.html.
[3]http://www.cse.cuhk.edu.hk/~jkzhu/dup_detect.html.



(a) TRECVID2003 Dataset (b) TRECVID2004 dataset

Fig. 12. Cumulative accuracy of similarity measure and features using Nearest Neighbor Ranking on the TRECVID2003 dataset (600 keyframes) and the TRECVID2004 dataset (7006 keyframes).

For each grid, the mean value is extracted as a feature point. The size of the GIST feature vector is 512.

5.2.2 *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.

5.2.3 *Local Binary Pattern (LBP).* The local binary pattern [Ojala et al. 1996] 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.

5.2.4 *Gabor Wavelets Texture.* To extract Gabor texture features, each image is first scaled to 64×64 pixels. The Gabor wavelet transform [Lades et al. 1993] 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.

5.2.5 *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 [Canny 1986] to obtain the edge map for computing the edge orientation histogram. The 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.

Thus, a 809-dimensional vector in total is used to represent all the global features for each keyframe in the datasets.

5.3 Experiment I: Ranking with NN

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 first evaluate the retrieval performance of the global features with nearest neighbor ranking.

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 12. 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. Note that we normalize the feature vectors into zero mean and unit variance in order to ensure the numerical stability in computing the cosine distance measure. As a result, we employ the L_1 norm as the similarity measure in all of the remaining experiments.

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 12, 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 results using GIST feature outperforms the other four methods. Also, both GIST feature and gird color moment significantly outperform the other three methods, especially in a large dataset like TRECVID2004 dataset. Based on the experimental results, we also analyze the reason for the effectiveness of GIST features in detail. In contrast to grid color moment and color histogram, GIST features are robust to large illumination changes and the occlusions due to video editing. Moreover, GIST features obtain good results on the NDK with cluttered background. From these observations, we can conclude that this mainly benefits from the effective Gabor filters built-in GIST method which can capture the important structure information in images.

5.4 Experiment II: Reranking with S^{3}VM

In this part, 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 [Zhang and Chang 2004] and color moments [Zhao et al. 2006]. 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 13 and Figure 15(b) show the experimental 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. To make a fair comparison, we also apply the S^{3}VM method on the color moment. It can be clearly observed that the semi-supervised ranking boosts the performance of the color moment method at a large margin. Compared with the unsupervised Nearest Neighbor ranking and supervised SVM ranking methods, S^{3}VM also obtains better results. In particular, S^{3}VM ranking method achieves very high cumulative accuracy with the top 30 returns, about 98.3%.

5.5 Experiment III: Reranking with NIM

5.5.1 *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. In contrast to the nonrigid surface detection in Zhu and Lyu [2007] and Zhu et al. [2009], we employ a relatively small regularization coefficient λ_r in order to allow large deformations, which is set to 5×10^{-5} in our experiments. 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.

5.5.2 *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

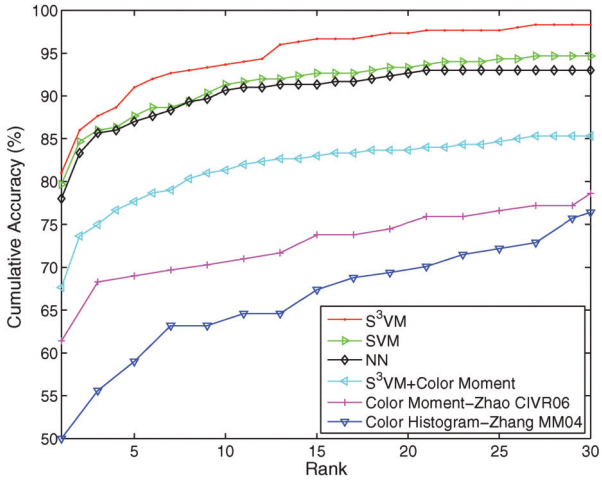
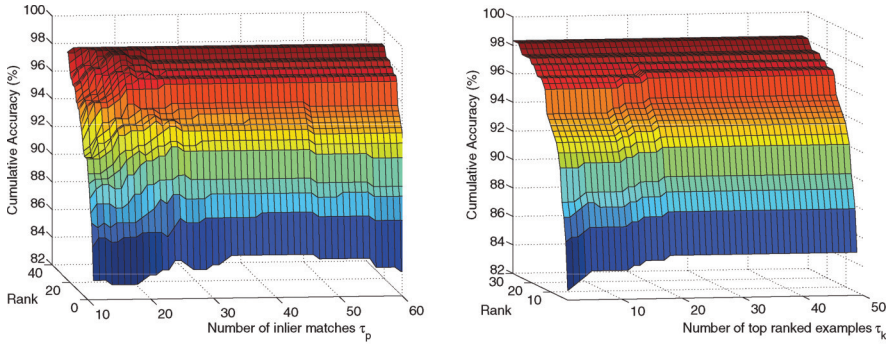


Fig. 13. Comparison of the proposed semi-supervised ranking method using S^{3}VM algorithm with other appearance based methods on the TRECVID2003 dataset.

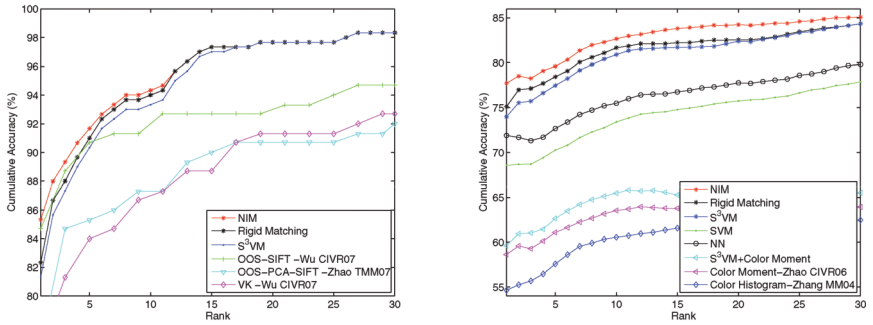


(a) different τ_p values (b) different τ_k values

Fig. 14. Cumulative accuracy of NDK retrieval using NIM method on the TRECVID2003 dataset. (a) A wide range of values can be chosen to determine the threshold τ_p. Image pairs of less than 30 inlier matches are viewed as nonduplicate in our experiments. (b) The overall accuracy grows with the number of top returned examples. We choose 50 as a trade-off between accuracy and computational cost.

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 does not seem too difficult. To justify this, we evaluate the performance by varying the τ_p values. Figure 14 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 25 and 40.

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 is, the more the computational cost incurrs. 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 14(b) shows the surface of cumulative accuracies with the top 30 returned



(a) TRECVID2003 dataset (b) TRECVID2004 dataset

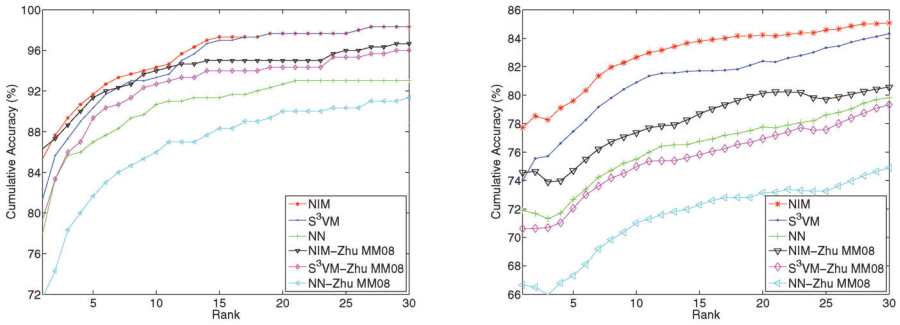
Fig. 15. Comparison of cumulative accuracy of NDK retrieval results on the TRECVID2003 dataset (600 images) and the TRECVID2004 dataset (7006 images), respectively.

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 40 to achieve good efficiency. We will evaluate the efficiency performance in a subsequent part of this article.

5.5.3 *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 [Wu et al. 2007c], the OOS-PCA-SIFT method [Zhao et al. 2007], and the Visual Keywords (VK) methods [Zhao et al. 2007]. Beside the qualitative comparison in Section 4.4, we also evaluate the NDK retrieval performance of the method based on the projective geometry assumption using homography.

For the TRECVID2003 dataset, it is relatively small and widely used as a benchmark testbed for NDK retrieval in literature. Figure 15(a) shows the experimental results of the cumulative accuracy of the top 30 returned keyframes on the TRECVID2003 dataset. From the experimental results, we can draw several observations. First of all, the proposed S^3VM method with global features outperforms the OOS-PCA-SIFT method [Zhao et al. 2007] and the VK method [Wu et al. 2007c], 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 improves the S^3VM ranking results at a large margin. In particular, NIM achieves more than 8% improvement on the rank one accuracy over S^3VM. Thus, among all compared methods, the proposed NIM method achieves the best performance, outperforming the state-of-the-art OOS-SIFT method [Wu et al. 2007c]. Also, NIM is more effective than the rigid matching method. Finally, the proposed NIM method achieves 85.3% top one retrieval result. This outperforms the most recent Spatially Aligned Pyramid Matching method [Xu et al. 2008] which reported 80.7% top one retrieval accuracy, by around 5.7%.

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 15(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. We can find that both NIM and rigid matching method improve the S^3VM ranking results. In particular, NIM outperforms the method based on rigid projective geometry assumption.



(a) TRECVID2003 Dataset (b) TRECVID2004 dataset

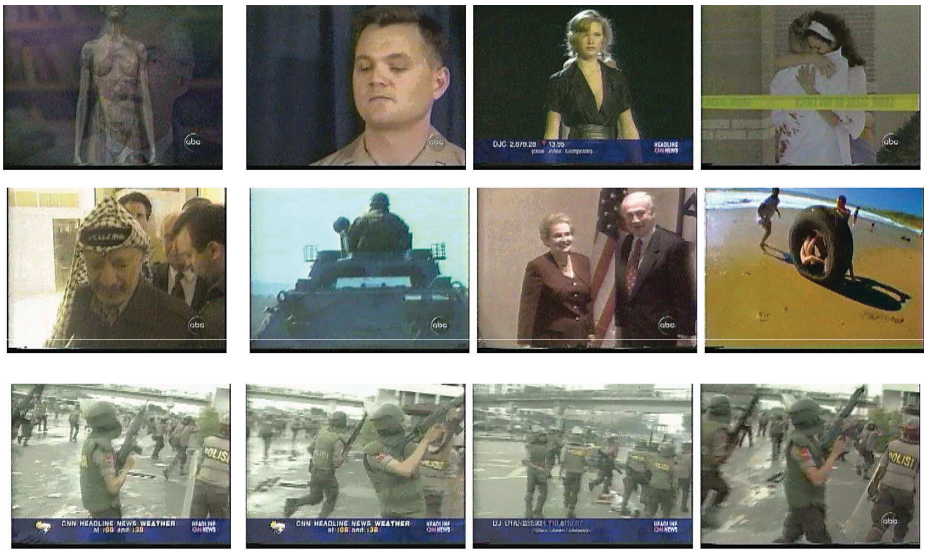
Fig. 16. Comparison of the proposed approach with the previous method on the TRECVID2003 dataset (600 images) and the TRECVID2004 dataset (7006 images), respectively.

It is also interesting to compare the proposed methods with our recent work in Zhu et al. [2008b]. We plot the NDK retrieval performance evaluation results on both TRECVID2003 dataset and TRECVID-2004 dataset in Figure 16. From the results, we first observe that the propose methods all outperform their corresponding algorithms in Zhu et al. [2008b] on both datasets. More specifically, the NN ranking method in this paper performs significantly better than the method without GIST feature. Specifically, there is around 10% improvement over the NN ranking in Zhu et al. [2008b]. Based on the improved NN ranking method, both S^3VM and NIM achieve better results, with around 5% improvement over the previous methods on TRECVID2004 dataset.

Finally, to give more insights on the proposed technique, we are interested in checking when our method may fail. To this purpose, we here briefly analyze the cases and attempt to find some possible reasons. Figure 17 shows some failure cases, where all of the top one retrieved examples by the proposed method are not the true duplicates. We can roughly categorize the failure cases into three groups. The main reason for the first group is that the query image is too blur or smooth to extract the discriminative feature points by the local feature detector. In particular, some transition video effects will lead to the blur keyframes, as shown in the first row. Note that keyframe extractors usually report the transition as keyframe. As for the second group, NIM method fails because that the spatial variations between NDK are too large to be modeled as nonrigid mapping, especially for the object movements in different directions. The second row shows an example. For the last group, the false results are mainly due to the visual similarities across NDK; for example crowds often share very similar global and local features. As shown in the third row, the first two examples are visually very similar to the query image, but they are not labeled as the true duplicates according to the ground truth. We also have some qualitative analysis of these failure cases on both TRECVID2003 and TRECVID2004 dataset. According to the above definition, we find the proportions of the failure cases in each category are 38%, 34%, and 28% on TRECVID2004 dataset and 26%, 35%, and 39% on TRECVID2004 dataset, respectively. This indicates that the proportion of failure cases in each category is similar, and visually similar, NDK leads to more false positives on a large dataset.

5.6 Evaluation of Computational Cost

Finally, we empirically study the efficiency performance of the proposed NIM and S^3VM methods. In our experiment, both the local features and global appearance features are extracted offline. Table I and Table II summarize the overall computational time for comparing all pairs of keyframes on both datasets. From these results, it can be observed that NIM is more efficient than the OOS-SIFT



(a) Query Image (b) Top three retrieval results

Fig. 17. Failure examples in showing when the presented method may fail.

Table I. Comparison of Overall Time Cost of 300 Queries on the TRECVID2003 Dataset

NIM	S^3VM	NN	OOS [Wu et al. 2007]	VK [Wu et al. 2007]
12.7min	3sec	1sec	6.5hour	1.5min

Table II. Comparison of Overall Time Cost of 1,953 Queries on the TRECVID2004 Dataset

NIM	S^3VM	NN	OOS [Wu et al. 2007]	VK [Wu et al. 2007]
83.5min	8.1min	30sec	N/A	N/A

method [Wu et al. 2007] and less efficient than the VK method, which simply computes the similarity of visual words. Note that the VK approach usually requires much preprocessing time for extracting the visual keywords offline. Additionally, we clearly see that the methods using global features are significantly more efficient than the ones using local feature matching. This again demonstrates the effectiveness and importance of the proposed multilevel ranking scheme for improving 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 18. 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 the real-world applications, one can choose an appropriate τ_k to balance the tradeoff between accuracy and efficiency. For example, when τ_k equals 10, each query for NIM takes about 1 second and achieves rather high cumulative accuracy, at about 93%.

6. CONCLUSIONS AND FUTURE WORK

In this article, a novel method is presented for dealing with the Near-Duplicate Keyframe (NDK) retrieval, which comprises an effective multilevel ranking scheme and a novel nonrigid image matching

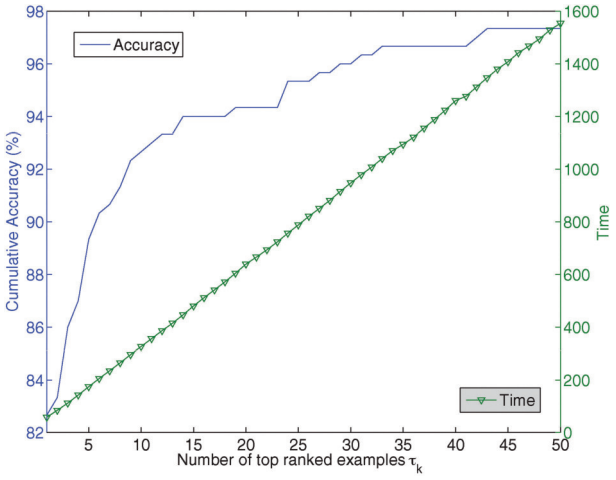


Fig. 18. Computational efficiency and retrieval performance on the TRECVID2003 dataset. The *left* vertical axis shows mean *cumulative accuracy* of the top 30 returned results, and the *right* vertical axis represents the overall time cost for all 300 queries.

technique. To reduce the overall computational cost, we also proposed an effective multilevel ranking scheme together with a semi-supervised ranking technique using semi-supervised SVM (S^3VM) to boost the ranking performance with the unlabeled data. Compared with conventional approaches with either bipartite graph matching or projective geometry, the proposed nonrigid image matching (NIM) algorithm not only recovers the explicit nonrigid mapping between two NDKs but also effectively locates the correct correspondences using a robust coarse-to-fine optimization scheme. Moreover, our method can detect the NDK pairs accurately while recovering the local deformations between them simultaneously. Furthermore, we also developed a new feature extraction method to improve the NDK retrieval accuracy. We conducted extensive evaluations on two testbeds extracted from the TRECVID corpora. The encouraging experimental results demonstrated that our approach is clearly more effective than traditional methods, especially in dealing with cases involving local deformations and viewpoint changes, which commonly occur in practice.

Despite promising results obtained by the proposed methods, some limitations and future directions should be mentioned. First of all, the feature representation scheme can be further improved. Our current approach simply concatenates various features, in which weights of different features are not fully optimized. Feature selection and feature weighting techniques can be studied to further enhance the retrieval performance. Second, our work focuses on the NDK retrieval task, whereas it may be more appropriate to detect NDKs directly from a collection of images in some applications. In contrast to the retrieval task, one challenge to the NDK detection task is to determine optimal thresholds for reporting NDKs. We will explore this in future work. Finally, the efficiency of the proposed multilevel ranking scheme can be further improved. In particular, the nearest neighbor ranking stage could be still computationally intensive for very large-scale datasets. In future work, we can improve this by adopting some effective high-dimensional indexing techniques, such as locality-sensitive hashing [Andoni and Indyk 2008].

ACKNOWLEDGMENTS

The authors would like to thank Prof. C.W. Ngo and Mr. W.L. Zhao for providing their experimental results. The authors also thank the reviewers and associate editor for their helpful comments.

REFERENCES

ANDONI, A. AND INDYK, P. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Comm. ACM 51,* 1, 117–122.

BAY, H., TUYTELAARS, T., AND GOOL, L. J. V. 2006. Surf: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*. 404–417.

BOYD, S. AND VANDENBERGHE, L. 2004. *Convex Optimization*. Cambridge University Press.

CANNY, J. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Anal. Mach. Intell. 8,* 6, 679–698.

CHUM, O. AND MATAS, J. 2005. Matching with prosac- progressive sample consensus. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. Vol. 1. 220–226.

CHUM, O., PHILBIN, J., ISARD, M., AND ZISSERMAN, A. 2007. Scalable near identical image and shot detection. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07)*. 549–556.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.citeulike.org/user/Comm.doubleshow/tag/file-import-09-04-17.

FISCHLER, M. A. AND BOLLES, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. CACM 24,* 6, 381–395.

FUA, P. AND LECLERC, Y. 1995. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int. J. Comput. Visi. 16,* 1, 35–56.

FUKUNAGA, K. 1990. *Introduction to Statistical Pattern Recognition*. Academic Press Professional, Inc.

HOI, C.-H., WANG, W., AND LYU, M. R. 2003. A novel scheme for video similarity detection. In *Proceedings of the International Conference on Image and Video Retrieval*. 373–382.

HOI, S. C. AND LYU, M. R. 2008. A multi-modal and multi-level ranking framework for content-based video retrieval. *IEEE Trans. Multimed. 10,* 4, 607–619.

KASS, M., WITKIN, A., AND TERZOPOULOS, D. 1988. Snakes: Active contour models. *Int. J. Comput. Visi. 1,* 4, 321–331.

KE, Y., SUKTHANKAR, R., AND HUSTON, L. 2004. Efficient near-duplicate detction and sub-image retrieval system. In *Proceedings of ACM MULTIMEDIA*. ACM, 869–876.

LADES, M., VORBRUGGEN, J. C., BUHMANN, J., LANGE, J., VON DER MALSBURG, C., WURTZ, R. P., AND KONEN, W. 1993. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput. 42,* 5, 300–311.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Visi. 60,* 2, 91–110.

MIKOLAJCZYK, K. AND SCHMID, C. 2005. A performance evaluation of local descriptors. *IEEE Trans. Patt. Analys. Mach. Intel. 27,* 10, 1615–1630.

NGO, C.-W., ZHAO, W.-L., AND JIANG, Y.-G. 2006. Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In *Proceedings of ACM MULTIMEDIA*. ACM, 845–854.

OJALA, T., PIETIKAINEN, M., AND HARWOOD, D. 1996. A comparative study of texture measures with classification based on feature distributions. *Patt. Recog. 29,* 1, 51–59.

OLIVA, A. AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Visi. 42,* 3, 145–175.

PILET, J., LEPETIT, V., AND FUA, P. 2008. Fast non-rigid surface detection, registration, and realistic augmentation. *Int. J. Comput. Visi. 76,* 2, 109–122.

QAMRA, A., MENG, Y., AND CHANG, E. Y. 2005. Enhanced perceptual distance functions and indexing for image replica recognition. *IEEE Trans. Patt. Anal. Mach. Intell. 27,* 3, 379–391.

RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Visi. 40,* 2, 99–121.

SINDHWANI, V., NIYOGI, P., AND BELKIN, M. 2005. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the International Conference on Machine Learning*. ACM Press, 824–831.

SIVIC, J. AND ZISSERMAN, A. 2003. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision (ICCV'3)*. 1470–1477.

SMEATON, A. F., OVER, P., AND KRAAIJ, W. 2006. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*. ACM Press, New York, NY, 321–330.

VAPNIK, V. N. 1998. *Statistical Learning Theory*. John Wiley & Sons.

WU, X., HAUPTMANN, A. G., AND NGO, C.-W. 2007a. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proceedings of ACM MULTIMEDIA*. ACM, 168–177.

WU, X., HAUPTMANN, A. G., AND NGO, C.-W. 2007b. Practical elimination of near-duplicates from web video search. In *Proceedings of ACM MULTIMEDIA*. ACM, 218–227.

WU, X., ZHAO, W.-L., AND NGO, C.-W. 2007c. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proceedings of the International Conference on Image and Video Retrieval*. ACM, 162–169.

XU, D., CHAM, T.-J., YAN, S., AND CHANG, S.-F. 2008. Near duplicate image identification with spatially aligned pyramid matching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

XU, Z., JIN, R., ZHU, J., KING, I., AND LYU, M. R. 2007. Efficient convex relaxation for transductive support vector machine. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 1641–1648.

YAN, R., HAUPTMANN, A. G., AND JIN, R. 2003. Negative pseudo-relevance feedback in content-based video retrieval. In *Proceedings of ACM MULTIMEDIA*. 343–346.

ZHANG, D.-Q. AND CHANG, S.-F. 2004. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of ACM MULTIMEDIA*. ACM, 877–884.

ZHAO, W., CHELLAPPA, R., PHILLIPS, P. J., AND ROSENFELD, A. 2003. Face recognition: A literature survey. *ACM Comput. Surv. 35,* 4, 399–458.

ZHAO, W., JIANG, Y., AND NGO, C. 2006. Keyframe retrieval by keypoints: Can point-to-point matching help? In *Proceedings of the International Conference on Image and Video Retrieval*. 72–81.

ZHAO, W.-L., NGO, C.-W., TAN, H. K., AND WU, X. 2007. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Trans. Multimed. 9,* 5, 1037–1048.

ZHU, J., HOI, S. C., AND LYU, M. R. 2008a. Face annotation by transductive kernel fisher discriminant. *IEEE Trans. Multimed. 10,* 1, 86–96.

ZHU, J., HOI, S. C., LYU, M. R., AND YAN, S. 2008b. Near-duplicate keyframe retrieval by nonrigid image matching. In *Proceedings of ACM MULTIMEDIA*. 41–50.

ZHU, J., HOI, S. C., XU, Z., AND LYU, M. R. 2008c. An effective approach to 3d deformable surface tracking. In *Proceedings of the European Conference on Computer Vision*. III: 766–779.

ZHU, J. AND LYU, M. R. 2007. Progressive finite newton approach to real-time nonrigid surface detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

ZHU, J., LYU, M. R., AND HUANG, T. S. 2009. A fast 2d shape recovery approach by fusing features and appearance. *IEEE Trans. Patt. Anal. Mach. Intell. 31,* 7, 1210–1224.

ZHU, X. 2005. Semi-supervised learning literature survey. Tech. rep., Carnegie Mellon University.

Received November 2008; revised May 2009, August 2009; accepted September 2009