Detecting Image Near-Duplicate by Stochastic Attribute Relational Graph Matching with Learning

Dong-Qing Zhang and Shih-Fu Chang

Department of Electrical Engineering,
Columbia University, New York, NY 10027.
{dqzhang,sfchang}@ee.columbia.edu

ABSTRACT

Detecting Image Near-Duplicate (IND) is an important problem in a variety of applications, such as copyright infringement detection and multimedia linking. Traditional image similarity models are often difficult to identify IND due to their inability to capture scene composition and semantics. We present a part-based image similarity measure derived from stochastic matching of Attribute Relational Graphs that represent the compositional parts and part relations of image scenes. Such a similarity model is fundamentally different from traditional approaches using low-level features or image alignment. The advantage of this model is its ability to accommodate spatial attribute relations and support supervised and unsupervised learning from training data. The experiments compare the presented model with several prior similarity models, such as color histogram, local edge descriptor, etc. The presented model outperforms the prior approaches with large margin.

Keywords

Image Near-Duplicate Detection, Attribute Relational Graph Matching, Part-based Image Similarity Measure

1. INTRODUCTION

Image Near-Duplicate (IND) refers to a pair of images in which one is close to the exact duplicate of the other, but different in the capturing conditions, times, rendering conditions, or editing operations. Detection and retrieval of IND is very useful in a variety of real-world applications. For example, in the context of copyright infringement detection, one can identify likely copyright violation by searching over the Internet for the unauthorized use of images. In multimedia content management, detecting image duplicates and near-duplicates can help link the stories in multi-source videos, articles in press and web pages. For example, in an environment that has broadcast news video from multiple channels, one important task is to thread the news

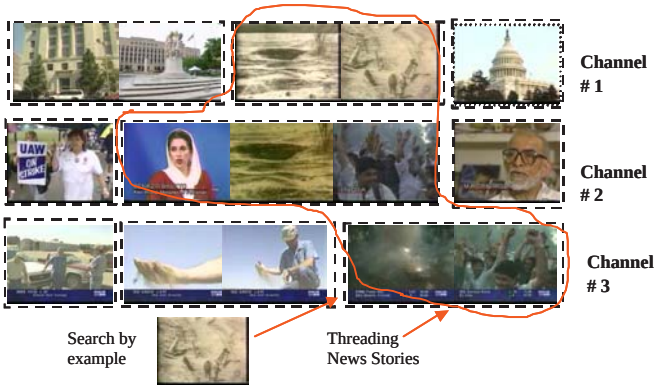

Figure 1: Threading news storied across channels

videos that come from different sources and countries to discover correlated topics and reconstruct semantics. We found that a significant number of news video stories contain near-duplicate images, which often provide strong clues for the topic-level similarities of news videos. In the TREC-VID 2003 corpus [1], out of the 362 news stories corresponding the predefined TDT2 news topics, there are 38 stories that contain non-anchor INDs. Figure 1 shows the scenario of news story threading, using the detected IND as the cue. Two images in IND may be captured at different times and/or by different cameras, resulting in significant scene variations, including occlusions and movements of objects, illumination changes, scene geometry difference, and color /contrast difference caused by different devices. Such variations bring about great challenges for the development of automatic IND detection methods. Conventional image matching methods using low level features (e.g., color and edge) or models for detecting primitive concepts (e.g., indoor, car, people) will likely fail because of the great variations within the IND class and sometimes the high similarity among images in the non-IND class. Furthermore, determining whether two images are near-duplicate or not is often subjective. This necessitates a learning-based framework for IND detection. Conventional computer vision based approaches, for example wide-baseline matching, registration, is not able to the similarity from training data.

This paper describes a part-based image similarity measure for accurate IND detection. An image scene is represented

by an Attribute Relational Graph (ARG) that models the compositional parts as well as their relations of image scenes. The similarity model is thereby derived from the likelihood ratio of the stochastic transformation process that transforms one ARG to the other. Such a stochastic similarity model provides a principled framework for computing similarity. More importantly, the stochastic model enables the similarity measure to be learned from training data in both supervised and unsupervised fashions. Intuitively, for matching two images, this stochastic model characterizes a scene transformation model that accommodates the variations of two near-duplicate images, including movements of objects, occlusions and appearances of new objects.

We show that the approximate computation of the likelihood ratio can be realized by the well-known Loopy Belief Propagation (LBP) algorithm. Fast LBP scheme is developed to speed up the likelihood computation. Learning of the similarity model is carried out in the vertex-level with supervised fashion and in the graph-level with unsupervised manner using Expectation-Maximization.

To the best of our knowledge, this is the first statistical framework that accommodates all types of variations of IND and the first part-based image similarity model supporting both supervised and unsupervised learning. The framework is also general in the sense that more features of the image can be easily incorporated.

The paper is organized as following: Section 2 presents the prior work on image duplicate detection, image similarity model and graph matching. Section 3 analyzes the variations of image near-duplicates in multimedia databases. Section 4. presents the part-based representation for image scene, followed by Section 5. on matching Attribute Relational Graph. Section 6. introduces camera model for enhancing the baseline system. Section 7. describes the experimental set-up and result analysis.

2. RELATED PRIOR WORK

Image Exact-Duplicate (IED) detection has been exploited in various contexts. In [3][**?**], image and video copy detection was investigated using image similarity measure by low-level features, for instance, color histogram and local edge descriptor. In contrast, rare work can be found on IND detection. The published work we can find is [7], the IND detection is realized by global camera calibration, change area detection and analysis of change area. The framework is established in a rule-based manner and does not support learning distance measure from training data.

Image similarity model has been studied for decades. Indexing by low-level features is an efficient technique, but lacks the discriminative power for IND detection due to its inability to capture scene semantics and composition. Semantic representation, such as the model vector system [11][9] could be a promising technique for IND detection. However, the limitations of the model vector system for IND detection are two folds: the vector representation of concepts fails to capture spatial/attribute relations of individual concepts; Most concept detectors require a large number of examples for learning.

Part-based image similarity has been previously pursued using 2-D string [4] and composite region templates (CRTs)[13]. However string representation is difficult to accurately represent visual scenes with complicated visual features, while region-based representation is sensitive to the errors of region segmentation. Part-based approaches have also been used in object/scene recognition [5] and scene generative model [8]. The approach in this paper differs from scene generative models in the sense that we do not assume scene models for individual scene categories.

Previous methods for attribute relational graph matching are realized by energy minimization [6], relaxation [10] and spectral method [12]. However, none of these prior work provide framework for supervised and unsupervised learning of the similarity measure from training data.

3. DEFINITION AND ANALYSIS OF IMAGE NEAR-DUPLICATE

The definition of Image Near-Duplicate has been previously provided by [7]. The following definition also considers IND generated by human editing and of computer generated images.

Definition 1. Image Near-Duplicate (IND) is a pair of images in which one is close to the exact duplicate of the other, but differs slightly due to variations of capturing conditions (camera, camera parameter, view angle, etc), acquisition time, rendering conditions, or editing operations.

The above definition is not completely unambiguous. For example, given two images taken from the same site at slightly different time, one tends to consider them IND. But such a decision may be ambiguous and greatly depends on the context. For example, images of a site taken hours (or even months) apart may have significant difference and will not be considered IND. However, in some scenarios such as broadcast news video, such ambiguity can be greatly reduced due to production conventions and typical user expectation. For example, IND showing visuals of the same site/event in breaking news are normally recognizable without much controversy.

In addition, we do not distinguish IND from exact duplicates, as the latter refers to exactly identical data, detection of which is not challenging. We aim at discrimination of IND from different images, which will be referred to as non-duplicate images.

3.1 Variations of Duplicates

The variations of two near-duplicate images can be caused by a few reasons, which can be categorized as: content change, camera change and digitization process. Figure 2. illustrates the causes of the variations of two near-duplicate images.

In photos and videos, the variations of image near-duplicates can be further categorized as follows[7]:

- Scene changes: movement and occlusion of foreground

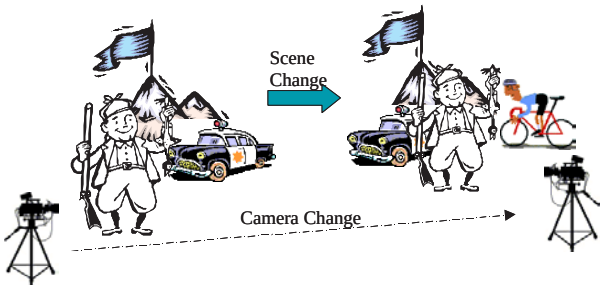

Figure 2: The generation of IND

objects, absence or presence of foreground objects, background change, post editing of images, etc.

- Camera parameter changes: camera view point change, camera tilting, panning, zooming, etc.
- Photometric changes: change of exposure or lighting condition, etc.
- Digitization changes: hue shift, contrast change,resolution change, etc.

Different image sources may have different types of duplicate variations. For example, for photos, variations are mostly caused by camera parameter changes, because the photographers tend to take more than one pictures with different camera settings so that they can select the best result later. In contrast, in broadcast news videos, variations are often caused by content change due to the time difference of capturing duplicate images. Figure 3 shows several IND examples illustrating a range of IND variations. The images are extracted from TREC-VID 2003 news video corpus.

Figure 3: Different variations of IND in News Video Databases (Top: content change, Mid: lighting change, Bottom: camera change

3.2 A Case Study of Duplicates in News Videos

To show the usefulness of IND detection in the scenario of news video analysis. We conducted a statistical study on the distribution and variations of INDs in the TREC-VID 2003 video corpus[1]. The TREC-VID2003 corpus consists of news videos from CNN headline and ABC world news from January 1998 to June 1998. The closed captions and speech transcripts of these videos were also used as the benchmark data for Topic Detection and Tracking (TDT) [2], a major evaluation event in the field of natural language processing. The TREC-VID2003 corpus is partitioned into two subsets, the development set and the test set. The development set consists of 2128 news stories according to the TDT2 story segmentation ground truth. These news stories are grouped by the TDT annotators into 28 topic clusters(Note not all stories are labeled in TDT2). The ground truth of the 28 clusters cover 362 stories in the corpus. Among these 362 stories, at least 38 video stories are found to contain duplicate or near-duplicate frames. These stories are distributed across 10 topics. Table 1 lists some examples of INDs with their related topics.

Table 1: Examples of Duplicates in News Videos

Topic	# of Stories having Duplicates
India nuclear test	CNN: 1, ABC: 3
Jerusalem	CNN: 1, ABC: 1
Indonesia violence	CNN: 1, ABC: 2
GM Strike	CNN: 4, ABC: 5
School Shooting	CNN: 1, ABC: 1

To investigate the variations of INDs, we extract 150 IND pairs from the TREC-VID 2003 video corpus. The following table manifests the variations of the extracted IND pairs. These IND images are also used as the benchmark data for performance evaluation.

Table 2: Breakout of INDs in news videos over different types of variations

Obj Move	New Obj	Hue
75	64	24
Illuminance	Camera Angle	Camera Zoom
52	40	34

3.3 IND Detection and Retrieval

Essentially, there are two types of problems related to IND: *IND retrieval*, aims at finding all images that are duplicate or near-duplicate to an input query image. The problem arises in the context of copyright violation detection, and query-by-example applications. *IND detection* aims at finding all duplicate image pairs given all possible pairs from the image source. The detected INDs could be used to link news stories and group them into threads.

IND detection is more difficult than *IND retrieval*. Because for the latter case, a query image has been chosen by a user, who usually has some belief that a near-duplicate image exists in the database. The retrieval task is considered successful if the target appears in the returned images within certain search scope.

The difficulty of *IND detection* is due to the fact that as the size of database increases, the number of duplicate pairs generally increases in a linear speed, while the number of non-duplicate pairs increases in a quadratic speed, leading to the deterioration of the detection precision.

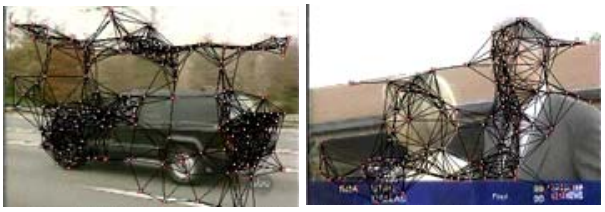

Figure 4: Interest point based ARGs

4. PART-BASED REPRESENTATION OF VISUAL SCENE

Part-based image representation aims at representing the visual scene by its compositional parts and relations of parts. The parts could be regions or interest points. Mathematically, part-based representation can be realized by Attribute Relational Graph (ARG), defined as following

Definition 2. An attribute relational graph is a triple $G = (V, E, A)$, where V is the vertex set , E is the edge set, and A is the attribute set that contains unary attribute a_i attaching to each node $n_i \in V$, and binary attribute a_{ij} attaching to each edge $e_k = (n_i, n_j) \in E$.

The unary attributes of an ARG characterize the properties of the individual parts, while the binary attributes signify the part relations. Parts can be captured by two popular fashions: using salient feature points, or using segmented regions. Region based representation is sensitive to region segmentation errors, therefore we use interest point based approach in our system. The Figure 4 shows the ARGs with interest point representation.

We detect the interest points by using the SUSAN corner detector[14]. Each vertex n_i of an ARG is attached with a 11-dimension feature vector y_i: two components for spatial and vertical coordinates; three components for RGB color; six components for Gabor wavelet features. The Gabor wavelet features are extracted using Gabor filter banks with two scales and three orientations (with window size 13x13 and 25x25). Only magnitudes of the filter responses are included in the feature vector, and phase features are discarded. Each edge (n_i, n_j) of an ARG is attached with a 2-dimension feature vector y_{ij} that represents the spatial coordinate differences of the two vertexes. Currently, fully-connected graphs are used to capture all possible pairwise relations of vertexes.

5. SIMILARITY MEASURE OF ARGS

In order to match two ARGs, we use a stochastic process to model the transformation from one ARG to the other. The similarity is measured by the likelihood ratio of the stochastic transformation process. We refer such a definition of data similarity as the *transformation principle of similarity*(TPS). For visual scenes, the referred stochastic process characterizes a scene transformation model accommodating the scene changes discussed above.

5.1 Stochastic ARG Matching

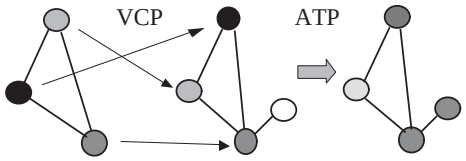

Figure 5: Stochastic Transformation Process for ARG similarity

Let H denotes the binary random variable corresponding to two hypotheses: $H = 1$, the target graph G_t is transformed from the source grough G_s, namely the two images are near-duplicate; $H = 0$, the two images are non-duplicate. Let $Y^s = \{\{Y_i^s\}; \{Y_{ij}^s\} | i, j \leq N\}$ denotes the unary and binary features of the source graph G_s; $Y^t = \{\{Y_u^t\}; \{Y_{uv}^t\} | u, v \leq M\}$ the features of the target graph G_s . Then we have the transform likelihood $p(Y^t|Y^s, H = 1)$, and $p(Y^t|Y^s, H = 0)$ respectively. The similarity between G_s and G_t is then defined as the likelihood ratio

$$S(G^s, G^t) = \frac{p(Y^t|Y^s, H = 1)}{p(Y^t|Y^s, H = 0)} \quad (1)$$

An image pair is classified as IND, if $S(G^s, G^t) > \lambda$. λ is a threshold. Adjusting λ yields different detection precision and recall. Note that in general $N \neq M$. We decompose the transformation process $p(Y^t|Y^s, H)$ into two steps. The first step copies the vertexes from G_s to G_t, referred to as *vertex copy process* (VCP). The second step transforms the attributes of the copied vertexes, referred to as *attribute transformation process*(ATP). This cascade stochastic process is visualized in Figure 5. The transformation process requires an intermediate variable to specify the correspondences between the vertexes in G_s and G_t. We denote it as X, referred to as *correspondence matrix*, a random 0-1 matrix taking value from $\chi = \{0, 1\}^{N \times M}$. $x_{iu} = 1$ means the ith vertex in G_s corresponds to the uth vertex in G_t. This is visualized in the Figure 6. For the case of one-to-one correspondences of vertexes in G_s and G_t, we need to have the following constraints

$$\sum_i x_{iu} \leq 1; \sum_u x_{iu} \leq 1 \quad (2)$$

By introducing X, the transformation process then can be factorized as following:

$$p(Y^t|Y^s, H) = \sum_{X \in \chi} p(Y^t|Y^s, X, H) p(X|Y^s, H) \quad (3)$$

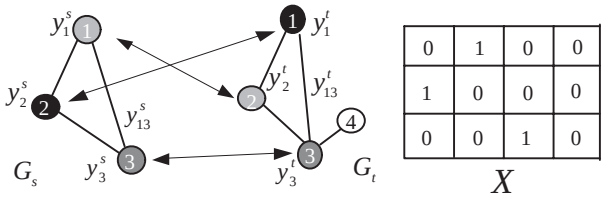

Figure 6: Vertex correspondences of two ARGs and the correspondence matrix

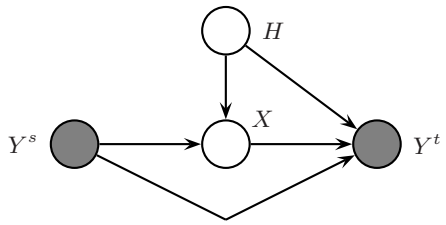

Figure 7: The generative model for ARG matching

$p(X|Y^s, H)$ characterizes the VCP and $p(Y^t|Y^s, X, H)$ represents the ATP. The transformation process is visualized as the generative graphical model in Figure 7. In order to satisfy the constraint Eq.(2), we let VCP be represented as a Markov network or Markov Random Field (MRF) with a two-node potential that ensures the one-to-one constraint being satisfied. The MRF model is written as

$$p(X|Y^s, H = h) = \frac{1}{Z(h)} \prod_{iu,jv} \psi_{iu,jv}(x_{iu}, x_{jv}) \prod_{iu} \phi_h(x_{iu})$$

where $Z(h)$ is the partition function. $h \in \{0, 1\}$ is the hypothesis. Note here we use notations iu and jv as the indices of the elements in the correspondence matrix X. For simplifying the model, we assume that X is independent on Y^s given H. The 2-node potential functions are defined as

$$\begin{aligned}&\psi_{iu,jv}(0,0) = \psi_{iu,jv}(0,1) = \psi_{iu,jv}(1,0) = 1, \quad \forall iu, jv\\ &\psi_{iu,jv}(1,1) = 0, \text{ for } i = j \text{ or } u = v;\\ &\psi_{iu,jv}(1,1) = 1, \text{ otherwise}\end{aligned}$$

The last 2 equations above enforce the one-to-one correspondence constraints listed in Eq.(2).The 1-node potential function is defined as

$$\phi_0(0) = p_0 \quad \phi_0(1) = q_0 \;; \quad \phi_1(0) = p_1 \quad \phi_1(1) = q_1$$

where p_0, q_0, p_1, q_1 controls the probability that the vertexes in the source graph are copied to the target graph under hypothesis $H = 0, H = 1$. Due to the partition function, any p_h, q_h with identical ratio p_h/q_h would result in the same distribution. Therefore, we can set p_h as 1, and let q_h be learned from training data.

It is not difficult to show that the partition function can be written as $Z(h) = \sum_{i=1}^{N} \binom{N}{i}\binom{M}{i} i! q_h^i$. For the efficient learning and likelihood calculation, we use the asymptotic approximation of $Z(h)$ as specified in the following theorem

THEOREM 1. *Let $N \leq M$. When $N \to \infty$, and $M - N < \infty$ The log partition function* $\log(Z(h))$ *tends to*

$$N\big[\log(q_h) + \lambda\big] \tag{4}$$

where λ is a constant written as

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \log\Big[\sum_{i=1}^{N} \binom{N}{i}\binom{N}{i} i!\Big]$$

which can be approximately calculated numerically.

Proof of this theorem can be found in our technical report [16].

For the *ATP*, we use naïve Bayes assumption to let all unary and binary features independent given Y^s and H. Therefore the *ATP* is fully factorized as

$$p(Y^t|X, Y^s, H) = \prod_{iu,jv} p(y_{uv}^t|x_{iu}, x_{jv}, y_{ij}^s) \prod_{iu} p(y_u^t|x_{iu}, y_i^s)$$

We assume that the Attribute Transformation Process is governed by Gaussian distributions with different parameters. Accordingly, we have conditional density functions for unary attributes

$$\begin{aligned}&p(y_u^t|x_{iu} = 1, y_i^s) = \mathcal{N}(y_i^s, \Sigma_1);\\ &p(y_u^t|x_{iu} = 0, y_i^s) = \mathcal{N}(y_i^s, \Sigma_0)\end{aligned} \tag{5}$$

And conditional density functions for binary attributes

$$\begin{aligned}&p(y_{uv}^t|x_{iu} = 1, x_{jv} = 1, y_{ij}^s) = \mathcal{N}(y_{ij}^s, \Sigma_{11})\\ &p(y_{uv}^t|(x_{iu} \cap x_{jv}) = 0, y_{ij}^s) = \mathcal{N}(y_{ij}^s, \Sigma_{00})\end{aligned} \tag{6}$$

The parameters Σ_1,Σ_{11},Σ_0,Σ_{00} are learned from training data.

5.2 Induced Markov Random Field

According to these setups, the term inside the equation (3) can be written as

$$\begin{aligned}&p(Y^t|Y^s, X, H = h)p(X|Y^s, H = h)\\ &= \frac{1}{Z(h)} \prod_{ij,uv} \psi'_{iu,jv}(x_{iu}, x_{jv}; y_{ij}^t, y_{uv}^s) \prod_{ij} \phi'_h(x_{iu}, y_i^s, y_u^t)\end{aligned}$$

Where

$$\begin{aligned}&\psi'_{iu,jv}(x_{iu}, x_{jv}; y_{ij}^s, y_{uv}^t) = \psi_{iu,jv}(x_{iu}, x_{jv})p(y_{uv}^t|x_{iu}, x_{jv}, y_{ij}^s)\\ &\phi'_h(x_{ij}, y_i^s, y_u^t) = \phi_h(x_{iu})p(y_u^t|x_{iu}, y_i^s)\end{aligned}$$

Therefore, the transformation likelihood becomes

$$p(Y^t|Y^s, H = h) = Z'(Y^t, Y^s, h)/Z(h) \tag{7}$$

where

$$Z'(Y^t, Y^s, h) = \sum_x \prod_{ij,uv} \psi'_{iu,jv}(x_{iu}, x_{jv}; y_{ij}^t, y_{uv}^s) \prod_{ij} \phi'_h(x_{iu}, y_i^s, y_u^t)$$

which in fact is the partition function of the following induced MRF, in which each of its vertexes corresponds to an entry of the correspondence matrix X

$$\begin{aligned}&p(X|Y^s, Y^t, H = h)\\ &= \frac{1}{Z'(Y^t, Y^s, h)} \prod_{ij,uv} \psi'_{iu,jv}(x_{ij}, x_{uv}; y_{ij}^t, y_{uv}^s) \prod_{ij} \phi'_h(x_{iu}, y_i^s, y_u^t)\end{aligned}$$

The partition function $Z(h)$ can be approximately computed from Eq.(4). The following section provides the approximate computation scheme for $Z'(Y^t, Y^s, h)$.

5.3 Computation of the Likelihood

The computation of the exact induced partition function $Z'(Y^t, Y^s, h)$ is intractable, since we have to sum over all possible X in the set χ, whose cardinality grows exponentially with respect to NM. Therefore we have to compute the likelihood using certain approximation.

5.3.1 Approximation of the Exact Likelihood

By applying the Jensen's inequality on the log partition $\log Z'(Y^t, Y^s, h)$, we can find its variational lower bound as following (more details can be found in [16]). The lower

bound then can be used as the approximation of the exact likelihood

$$\log Z' \geq \sum_{iu,jv} \hat{q}(x_{iu}, x_{jv}) \log \psi'_{iu,jv}(x_{iu}, x_{jv}; y^s_{ij}, y^t_{uv}) + \sum_{iu} \hat{q}(x_{iu}) \log \phi'_h(x_{iu}, y^s_i, y^t_u) + \mathcal{H}(\hat{q}(X)) \quad (8)$$

where $\hat{q}(x_{iu})$ and $\hat{q}(x_{iu}, x_{jv})$ are approximate one-node and two-node marginals, also known as beliefs. $\mathcal{H}(\hat{q}(X))$ is the entropy of $\hat{q}(X)$, which does not have tractable decomposition for loopy graphical model. $\mathcal{H}(\hat{q}(X))$ can be approximated using Bethe/Kikuchi approximation [15], by which [15] has shown that the $\hat{q}(x_{iu})$ and $\hat{q}(x_{iu}, x_{jv})$ can be obtained by a set of fixed point equations, known as Loopy Belief Propagation. The equation of the approximated entropy can be found in [15].

5.3.2 Fast Loopy Belief Propagation

The speed bottleneck of this algorithm is the Loopy Belief Propagation. Without any speedup, the complexity of BP is $O\big((N \times M)^3\big)$, which is formidable for computation. However, for complete graph, the complexity of BP can be reduced by introducing an auxiliary variable. The BP message update rule is modified as follows

$$m^{(t+1)}_{iu,jv}(x_{jv}) = k \sum_{x_{iu}} \psi_{iu,jv}(x_{iu}, x_{jv}) \phi_h(x_{iu}) M^{(t)}_{iu}(x_{iu}) / m^{(t)}_{jv,iu}(x_{iu})$$

$$M^{(t+1)}_{iu}(x_{iu}) = \exp\left(\Sigma_{kw \neq iu} \log(m^{(t+1)}_{kw,iu}(x_{iu}))\right)$$

Where M_{iu} is the introduced auxiliary variable, t is the iteration index. This modification results in $O\big((N \times M)^2\big)$ computation complexity.

To further speed up the likelihood computation, early determination of the node correspondences (i.e.x_{iu}) can be realized by thresholding the one-node probability in the Eq. (5). For example, if the coordinate of a vertex i in the source graph deviate too much from the coordinate of the vertex u in the target graph, the state variable x_{iu} can be fixed to zero with probability one. This operation is equivalent to reduce the vertex number of the induced MRF model, resulting in less computation cost.

5.4 Learning the Parameters

Learning ARG matching can be performed at two levels: vertex-level and graph-level. For the vertex level, the annotators annotate the correspondence of every vertex pair. This process is very expensive, since typically one image includes 50-200 interest points. In order to reduce human supervision, graph-level learning can be used, where annotators only indicate whether two images are IND or not without identifying specific corresponding interest points .

For learning at the vertex-level, maximum likelihood estimation results in the direct calculation of the mean and variance of Gaussian functions in the Eq.(5)(6) from the training data. The resulting estimates of the parameters are thereby used as the initial parameters for the graph-level learning.

For learning at the graph-level, we can use the following variational Expectation-Maximization (E-M) scheme:

E Step: Compute $\hat{q}(x_{iu})$ and $\hat{q}(x_{iu}, x_{jv})$ using Loopy Belief Propagation.

M Step: Maximize the lower bound in Eq.(8) by varying parameters. This can be realized by differentiating the lower bound with respect to the parameters, resulting in the following update equations

$$\begin{aligned}
\xi_{iu} &= \hat{q}(x_{iu} = 1); \quad \xi_{iu,jv} = \hat{q}(x_{iu} = 1, x_{jv} = 1) \\
\Sigma_1 &= \frac{\sum_k \sum_{iu} (y^t_u - y^s_i)(y^t_u - y^s_i)^T \xi_{iu}}{\sum_k \sum_{iu} \xi_{iu}} \\
\Sigma_0 &= \frac{\sum_k \sum_{iu} (y^t_u - y^s_i)(y^t_u - y^s_i)^T (1 - \xi_{iu})}{\sum_k \sum_{iu} (1 - \xi_{iu})} \\
\Sigma_{11} &= \frac{\sum_k \sum_{iu,jv} (y^t_{uv} - y^s_{ij})(y^t_{uv} - y^s_{ij})^T \xi_{iu,jv}}{\sum_k \sum_{iu,jv} \xi_{iu,jv}} \\
\Sigma_{00} &= \frac{\sum_k \sum_{iu,jv} (y^t_{uv} - y^s_{ij})(y^t_{uv} - y^s_{ij})^T (1 - \xi_{iu,jv})}{\sum_k \sum_{iu,jv} (1 - \xi_{iu,jv})}
\end{aligned}$$

Where k is the index for the instances of the training graph pairs.

For the prior parameter q_h, the larger ratio q_1/q_0 would result in the more contribution of the prior model to the overall transformation likelihood ratio. Currently, we use a gradient descent scheme to gradually increase the value of q_1 (start from 1) until we achieve the optimal discrimination of IND classification in the training data. More details can be found in [16].

6. CAMERA TRANSFORMATION

Camera change may result in significant interest point movements, resulting in low matching likelihood. Therefore, we exploit to use camera parameter estimation and calibration to avoid the mismatch if two scenes are nearly identical but are captured by different camera settings.

Two images captured by different camera parameters are coupled by multi-view geometry constraints. There are essentially two types of constraints: homography and fundamental matrix. Basically, homography transform captures the viewing direction change of the camera, while fundamental matrix captures other types of camera parameter changes. In the current project, we only implemented the homography transform.

Under homography constraint, one interest point in the source image, described by the homogenous coordinate vector $X_1 = (Tx_1, Ty_1, T)$ is mapped to one point $X_2 = (T'x_2, T'y_2, T')$ in the target image by the homography transform H:

$$X_2 = HX_1 \quad (9)$$

Here we use the vertex matching results yielded by the Loopy Belief Propagation. That is, if $x_{iu} > 0.5$, then we consider the interest point i and u correspondent. After the estimation of the homography, the spatial attributes in the source ARG are calibrated by applying the estimated homography matrix.

7. EXPERIMENTS AND RESULTS

The benchmark data are collected from the TREC-VID 2003 corpus. The data set consists of 150 IND pairs (300 images) and 300 non-duplicate images. 100 IND pairs and all 300 non-duplicate images are from the keyframe set provided by the TREC-VID 2003, while the remaining 50 IND pairs are non-keyframes. The entire data set is partitioned into the

training set and the testing set. The training set consists of 30 IND pairs and 60 non-duplicate images.

7.1 Implementation Issues

From the approximate likelihood equations, we observe that the likelihood ratio is not invariant to the sizes of the input graphs. To reduce the sensitivity of the size variation. We use the averaged likelihood ratio $S(G^s, G^t)/(MN)$ instead of $S(G^s, G^t)$. Furthermore, we assume that under negative hypothesis $H = 0$, there is no vertex correspondence between G_s and G_t. Therefore, all x_{iu} and $x_{iu,jv}$ are set to 0 with probability 1. And the Gaussian parameters for $H = 0$ are the same as those for $H = 1$. To reduce the computational cost, we use the early determination scheme by a thresholding process. As a result, the maximum size of the induced MRF is 150 vertexes. The average computation time of matching two images is about 0.4 second. The speed may be further boosted using more sophisticated fast BP schemes, as recently studied by some researchers.

7.2 Learning Procedure

Learning process consists of two phases. In the first phase, we apply the supervised vertex-level learning, where the correspondences of interest points are marked by the annotator. The number of interest points varies from 50 to 200 per image. Only 5 near-duplicate pairs and 5 non-duplicate pairs are used in vertex-level learning. In the second phase, we conduct the graph-level learning. E-M scheme is carried out using 25 IND pairs and 25 Non-IND pairs.

7.3 Performance Comparison with Previous Methods

The performance of the developed method (GRAPH) is compared with color histogram (CH), local edge descriptor (LED), averaged features distance of interest points (AFDIP) and graph matching with manual parameter setting (GRAPH-M). Local edge descriptor has been demonstrated as the best feature for image copy detection in the previous work [**?**]. For color histogram, we use HSV color space with 64 bins of H, 8 bins of S, and 4 bins of V. AFDIP is the summation of all possible cosine distances between the unary feature vectors of the interest points in G_s and G_t divided by NM. GRAPH-M is the graph matching likelihood ratio under the manually selected Guassian parameters. The parameters are obtained by manually adjusting the covariance matrices until we observe the best interest point matching results (by binarizing the belief $\hat{q}(x_{iu})$). The parameters are initialized using vertex-level learning with two IND pairs. *recall* is defined as the number of the correctly detected IND pairs divided by the number of all ground truth IND pairs. *Precision* is defined as the number of the correctly detected IND pairs divided by the number of all detected IND pairs.

From the results, we observe that the averaged feature distance performs even worse than color histogram while graph matching with training performs best. This indicates that the performance of graph matching mainly gains from the probabilistic inference and learning. Graph matching with learning also outperforms graph matching with manual parameter setting. This confirms that the learning process not only reduces the human labor cost but also significantly improves the performance by capturing data statistics.

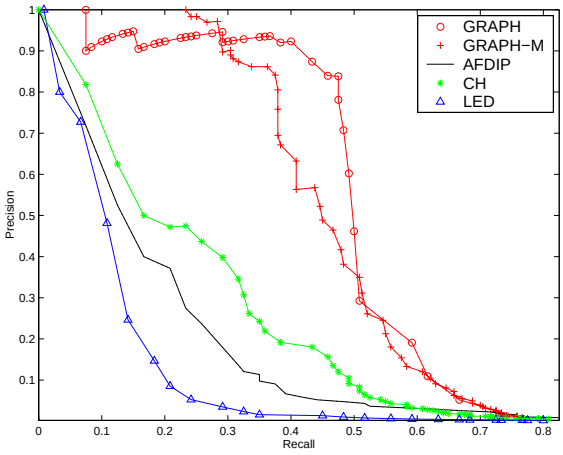

Figure 8: The performance comparison with the previous approaches and ARG matching without inference and learning

7.4 Sensitivity to Interest Point Detection

We study if the algorithm is sensitive to the interest point detection. The algorithm is applied with different parameters of the corner detection algorithm (GRAPH, GRAPH-L, GRAPH-S), resulting in different numbers of the interest points and slight changes of their locations. The Figure 9 shows that the sensitivity to the interest point detection is fairly low comparing with the performance variation across different methodologies. This demonstrates the robustness of our method by using interest point based representation.

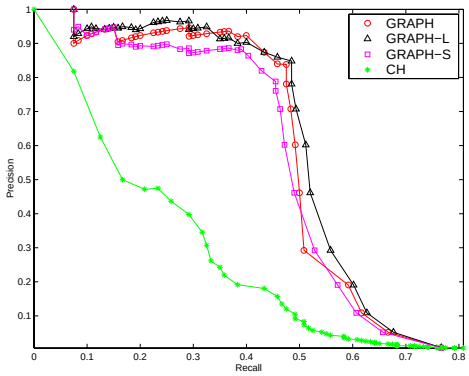

Figure 9: Study of sensitivity to different interest point detection parameters

7.5 Does Camera Transformation Help?

To see if the inclusion of the camera transform model indeed helps IND detection, we compare the performance of the system with camera transform and without camera transform. The results show that camera transform only contributes little performance improvement. This is likely due to inaccurate camera parameter estimation, since most INDs involve substantial content change. This again confirms that detecting IND in such a domain is challenging for the conventional image registration approaches.

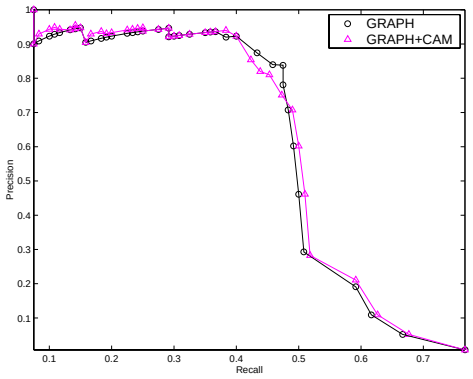

Figure 10: Study the performance gain by the inclusion of camera model

7.6 Duplicate Retrieval

Retrieval problem is important for some applications, such as copyright violation detection. Let Q_c denotes the number of queries that correctly retrieve its duplicate version in the top K retrieved images. Let Q denotes the overall number of queries. Then the retrieval performance is defined as $P(K) = Q_c/Q$, the probability of the successful top-K retrieval. In the current experiment, for each query image, we know there is one and only one duplicate version in the database.

As shown in the Figure 11, the graph matching approach outperforms others when $K \leq 50$. However, when K is larger than 50 (although in practice, users typically do not view more than 50 results), the performance of graph matching is lower than that of the color histogram. This is primarily because the likelihood ratio of graph matching quickly decays to nearly zero when two images are not matched. As a result, the ranks of these likelihood ratios become inaccurate. We also observed that the color histogram retrieved 90% duplicates when $K \geq 100$. This indicates that low cost approaches like color histogram can be used as a prefiltering step to narrow down the search scope, in which the more accurate graph matching method can be applied.

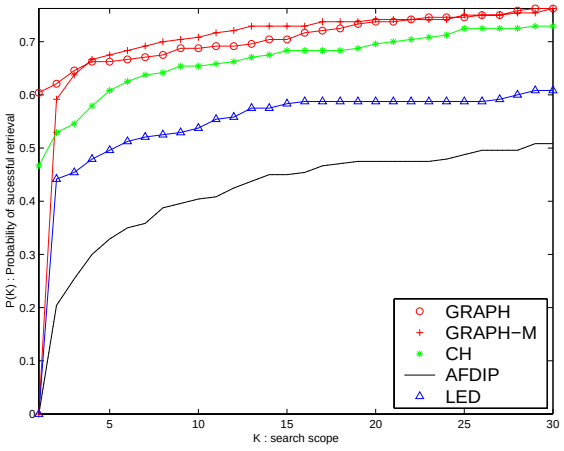

Figure 11: IND retrieval performance

8. CONCLUSIONS

We present a novel part-based image similarity measure by a learning-based matching of attribute relational graphs that represent images by their parts and part relations. Such a similarity model is fundamentally different from the conventional ones using low-level features or registrations. We show that the presented similarity measure outperforms previous approaches with large margin in detecting Image Near-Duplicate. The experiments also show that the similarity measure is insensitive to the parameter settings of interest point detection, and camera models contribute little performance improvement.

9. REFERENCES

[1] http://www-nlpir.nist.gov/projects/trecvid/. *Web site*, 2004.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.*, 1998.

[3] E. Chang, J.Wang, C.Li, and G. Wiederhold. Rime: A replicated image detector for the world wide web. In *Proceedings of SPIE Mulimedia Storage and Archiving Systems III*. IEEE, Nov. 1998.

[4] S.-K. Chang. *Principles of pictorial information systems design*. Prentice-Hall, Inc., 1989.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, pages 66–73. IEEE, 2003.

[6] H.G.Barrow and R. Popplestone. Relational descriptions in picture processing. *Machine Intelligence*, 6:377–396, 1971.

[7] A. Jaimes. Conceptual structures and computational methods for indexing and organization of visual information. *Ph.D. Thesis , Department of Electrical Engineering, Columbia University*, February 2003.

[8] N. Jojic, N. Petrovic, and T. Huang. Scene generative models for adaptive video fast forward. In *International Conference on Image Processing (ICIP), Barcelona, Spain, 2003*. IEEE, 2003.

[9] C.-Y. Lin, B. L. Tseng, M. Naphade, A. Natsev, and J. R. Smith. Videoal: A novel end-to-end mpeg-7 automatic labeling system. In *IEEE Intl. Conf. on Image Processing (ICIP)*. IEEE, September 2003.

[10] R.C.Wilson and E.R.Hancock. Structual matching by discrete relaxation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:1–2, 1997.

[11] J. R.Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *IEEE Intl. Conf. on Multimedia and Expo (ICME)*. IEEE, 2003.

[12] J. Shi and J. Malik. Self inducing relational distance and its application to image segmentation. *Lecture Notes in Computer Science*, 1406:528, 1998.

[13] J. R. Smith and C.-S. Li. Image classi cation and querying using composite region templates. *Journal of Computer Vision and Image Understanding*, 2000.

[14] S. Smith. A new class of corner finder. In *Proc. 3rd British Machine Vision Conference*, pages 139–148, 1992.

[15] J. S. Yedidia and W. T. Freeman. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, pages Chap. 8, pp. 239–236, January 2003.

[16] D.-Q. Zhang and S.-F. Chang. Stochastic attribute relational graph matching for image near-duplicate detection. *DVMM Techinical Report, Dept. of E.E., Columbia University*, July 2004.