Searching Visual Instances with Topology Checking and Context Modeling

Wei Zhang, Chong-Wah Ngo
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
wzhang34@student.cityu.edu.hk, cscwngo@cityu.edu.hk

ABSTRACT

Instance Search (INS) is a realistic problem initiated by TRECVID, which is to retrieve all occurrences of the querying object, location, or person from a large video collection. It is a fundamental problem with many applications, and also a challenging problem different from the traditional concept or near-duplicate (ND) search, since the relevancy is defined at instance level. True responses could exhibit various visual variations, such as being small on the image with different background, or showing a non-homography spatial configuration. Based on the Bag-of-Words model, we propose two techniques tailored for Instance Search. Specifically, we explore the use of (1) an elastic spatial topology checking technique based on Delaunay Triangulation (DT), and (2) a practical background context modeling method by simulating the "stare" behavior of human eyes. With DT, we improve the quality of visual matching by accumulating evidence from local topology-preserving patches, significantly boosting the ranks of topology consistent results. On the other hand, we increase the information quantity for visual matching with the "stare" model, such that instances appearing in both similar and different background can be highly ranked as results. The proposed techniques are evaluated on the INS datasets of TRECVID, achieving large performance gain with small computation overhead, compared with several existing methods.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Performance, Experimentation

Keywords

Instance Search; TRECVID; Spatial Topology Checking; Context Modeling

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.

1. INTRODUCTION

This paper addresses a practical problem in real life: given several visual examples of an instance topic, retrieve all video clips that contain the instance from a video dataset. With the increasing number of videos generated every day, searching for a certain topic (e.g., an object, a location, or a person) in large video collections [17] is a feature highly demanded by many applications, such as archive video search, personal video organization/browsing, law enforcement, protection of brand/logo. As shown in Figure 1, the problem of retrieving instances is challenging, considering the large visual variations introduced by totally different background (1st row), different viewpoints of 3D objects (2nd row), scale changes (3rd row), and small objects (1st and 4th rows).

The formal definition of Instance Search (INS) is initiated by TRECVID [17]: given several visual examples of a search topic with the corresponding binary masks indicating the locations of the instance, find all video segments that contain one or more occurrences of the query instance. Though similar, the problem of INS is different from its close relatives: concept-based search and Near-Duplicate (ND) search [4, 19]. For concept-based search, the relevancy is defined at the semantic level, and any results with the same semantic meaning meet the searching criteria. For example, searching "plane" should return planes with any types, colors, and sizes, while INS should only return the same plane as specified in the query. It also differs from ND search, where certain image operations (e.g., scaling, rotation, cropping, noises, and text overlay) are applied on the source image to produce NDs. For INS, the instance could appear in a totally different background context with different viewpoints, as long as the video segments contain the same instance. In general, ND search is more useful for whole image search, since both the instance and background can be exploited for visual matching. On the other hand, INS has less information to leverage, since the background is not necessarily useful in this case.

Although INS can be formulated as a traditional image retrieval problem, it has its own peculiarities in several ways:

1. Different from ND search, instances often occupy a small area on the image, and the background context is often different among images with the same instance.
2. A manually labeled ROI (Region-of-Interest) is often available for the query, so that we can distinguish the instance under query and background context. The labeling of ROI can be easily done with the help of touch screens.

Considering the characteristics of INS, a potential diffi-

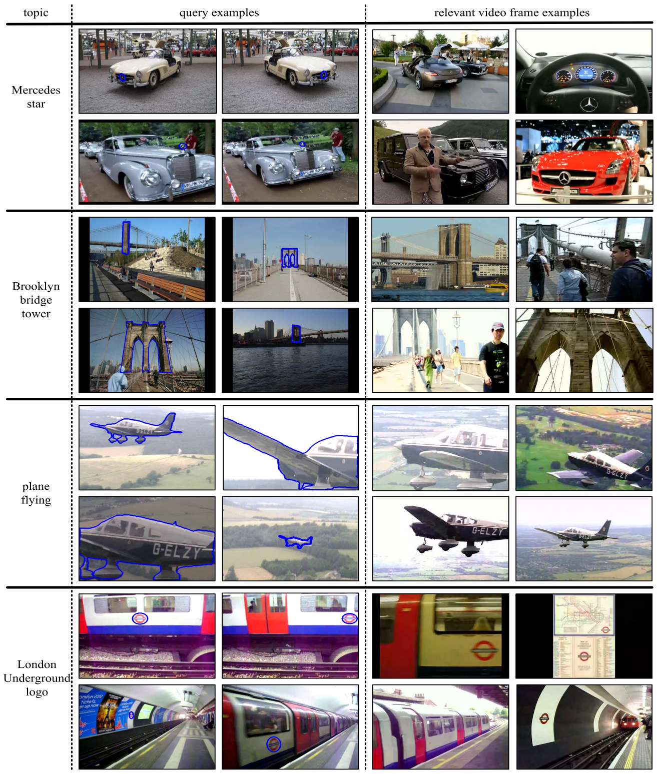

Figure 1: Examples of query topics in TRECVID datasets. Columns from left to right: topic name, image examples for the topic, and examples for relevant video keyframes. Contours of the ROI/mask are outlined with blue curves on each query image.

culty is that there is much less visual information to be exploited, especially when the size of query instance is small, as shown in Figure 1. Querying with such less information is error-prone, since the limited information is often not discriminative enough to retrieve instances out of a large video collection. As a result, lacking of quality and quantity information for visual matching becomes the major difficulty for INS. To tackle this problem, we propose two methods from different perspectives. The first one is by taking better use of the available information by modeling the spatial topology consistency. In other words, the lack of quantitative information is compensated by quality matching via topology checking. With an elastic topology model for spatial checking, we attempt to boost the ranking of relevant instances by making better use of the spatial information. Different from previous methods that impose a linear transformation over the absolute matching locations, we sketch and match the spatial topology based on Delaunay Triangulation. The second strategy seeks a way to increase the amount of query information by carefully considering the information from background context. Generally, we trust the information from the ROI, and it is risky to consider the areas outside. However, the information outside the ROI may enrich the limited information and provide more cues about the instance. The key is when to consider the context and how to weight the contribution from each parts. Inspired by the way of visual perception for human eyes, we enrich the instance under query by introducing a "stare" model to improve the discriminative power of instances.

The main contributions of this paper can be summarized as follows:

- We apply a triangulation based spatial consistency checking method originally proposed for image search [21]. The method emphasizes the topological consistency for quality matching. Rather than imposing a strict transformation for geometric consistency checking, a graph is constructed to encode the topology information for matched points. This gives better tolerance to true responses in INS by accumulating evidence from local regularities of the instance.
- Background context is modeled into the query by using the "stare" model that simulates the visual perception behavior of human eyes. Both NDs, which share common background with the query, and instances with novel background can be retrieved.

The remaining paper is organized as follows. Section 2 describes related work. Sections 3 presents the topology checking method with Delaunay Triangulation to improve the quality on visual matching, while Section 4 proposes the "stare" model for context modeling to increase the quantity of matchings. Section 5 presents our experimental results on the datasets of TV11 and TV12, and finally Section 6 concludes this paper.

2. RELATED WORK

The proposed method is rooted in the Bag-of-Words (BoW) retrieval model, which was initially used in text retrieval. Since introduced in [16], it has been widely used in multimedia retrieval community for its good tradeoff between performance and scalability. Standard BoW technique consists of several key components: image description, visual vocabulary, feature quantization, and inverted file. Images are first scanned for stable and representative regions [12, 11] and the local features [11] are extracted afterwards. The offline trained visual vocabulary defines a quantization function of the feature space, such that features quantized to the same visual word are considered to be similar. At the time of online retrieval, only a small subset of features is traversed with the help of inverted file.

Various modifications have been proposed to improve the original BoW method. Studies on fast vocabulary training [14] and the hierarchical structure [13] enable the million scale vocabulary with fine quantization [24]. Hamming Embedding [8], product quantization [10], and soft/multiple assignments [15, 8] further reduce the quantization error by better partitioning the feature space or smoothing the error. Query expansion [5] improves the recall significantly by formulating a refined query in each iteration.

Among all the variants of BoW, spatial consistency checking has always been a big branch, since the original BoW model makes no guarantee on the spatial regularity of visually matched patches. Filtering isolated matching points [16], bundling spatially-clustered features [20] impose a weak spatial constraint on visually matched feature points. However, these constraints are often too loose to reject the large number of false positives, especially for a large dataset. On the other hand, most of the strong spatial checking techniques are rooted in the planar homography [7] that requires the scene under view to be planar or the camera centers being at a fixed location. In practice, it is often approximated as the affine model [14], which works well for small areas, buildings with planar facades, or small viewpoint changes. Latest techniques rooted in the homography model include WGC [8], E-WGC [23], and GVP [22], which have gained great success in the past few years in terms of retrieval performance. However, the homography model works best on ND [19] or near identical [4] dataset. For 3D objects taken from different viewpoints, the exact spatial configuration is the epipolar geometry [7] with the fundamental matrix projecting a point to its epipolar line. Only a few works [2, 3] explore this model, which successfully retrieve some of the missed matching points for 3D structures. However, unlike the one-to-one mapping for homography, the fundamental matrix can only project a point on one image to a line on the other image, which is also considered to be a weak constraint. However, spatial checking for INS, which consists not only non-planar/3D structures (violating the planar homography), but also non-rigid instances (violating even the epipolar geometry), is seldomly addressed before.

The use of background context is trivial for traditional ND search, and most existing works just uses or discards the context completely. For example, since the background is also part of the query for ND search, it is fully included for retrieving ND images in [19, 4, 23]. For mobile search, where a ROI comes in handy, such as Google Goggles [1] and Snaptell[2], the background context is often ignored. However, for the problem of INS, the context information should be modeled to enrich the limited query information, in order to retrieve novel instances in different background as well as NDs when available.

3. SPATIAL TOPOLOGY CHECKING

While suitable for the visual similarity measurement, BoW does not guarantee the spatial regularity of matched features. However, spatial information is crucial for visual recognition, and is even important for instance retrieval, considering the limited number of features on the query target. Different from ND search, there exists lots of non-planar structures and 3D objects (Figure 1) that do not follow the planar homography transformation. Moreover, there are even non-rigid instances (e.g., person) that do not follow the epipolar geometry. Most of the existing works impose a linear transformation, which works best for planar and rigid instances. For general instance types including non-rigid and non-planar objects, we elastically model the spatial topology with Delaunay Triangulation based visual words matching [21]. In this section, we first summarize the essential spatial configurations [7] of corresponding points for INS, and then propose our spatial topology checking method.

3.1 Spatial Configurations for INS

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the homogeneous coordinates of corresponding points on the query and reference image, respectively. Spatial locations for planar scenes can be related by a planar homography matrix $\mathbf{H}$: $\mathbf{x}_1 = \mathbf{H}\mathbf{x}_2$, which defines a point-to-point mapping between correspondences of a planar plane or two views taken with fixed camera centers, which suits well for ND search. When it comes to 3D objects, the intrinsic projective geometry between two views becomes the epipolar geometry encapsulated by the fundamental matrix $\mathbf{F}$: $\mathbf{x}_1^{\mathbf{T}}\mathbf{F}\mathbf{x}_2 = \mathbf{0}$, which only defines a weak point-to-line mapping. Things are more complicated for INS, which includes plenty of 3D structures and even non-rigid instances (such as persons, animals). Neither planar homography nor epipolar geometry fits in this case. For example, previous spatial checking techniques fail on the the butterfly example in Figure 2, since the spatial variation is caused by a non-planar instance with a non-rigid motion. This section partially addresses this problem by proposing an elastic model that emphasizes the spatial topology regularity.

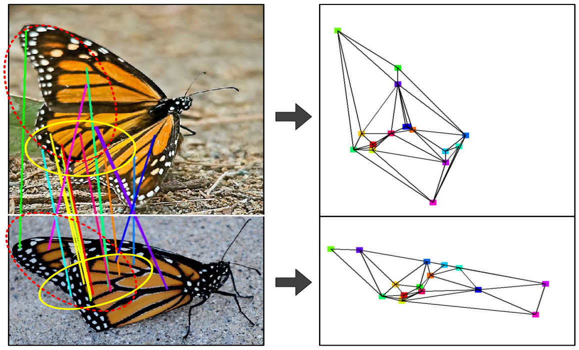

Figure 2: Illustration of the spatial topology checking method based on Delaunay Triangulation (DT). Left: two images with their matched features lined up. Note the matched words are indicated with the same color. Right: the triangulation graphs sketching the topology of matching points on the left.

3.2 Topology Checking via Triangulation

Delaunay Triangulation: This is a technique widely used in Computer Graphics for building meshes out of a set of points. A Delaunay Triangulation [6] for a set of points $\mathbf{P}$ on a 2D plane is a triangulation DT($\mathbf{P}$), so no point in $\mathbf{P}$ is inside the circumcircle of any triangle in DT($\mathbf{P}$). DT maximizes the sum of the minimum angles of all triangles after triangulation, such that regular/balanced triangles, rather than skinny ones, are preferred. Compared to other triangulation of the points, the smallest angle in Delaunay Triangulation is at least as large as the smallest angle in any other triangulations [1].
Motivation: Since there is no uniform transformation for the non-planar/non-rigid objects that occurs in INS, we seek for solutions from another perspective, i.e., topology. Our motivation is that the spatial topology tends to be stable for (1) different views of 3D objects for small/moderate viewpoint changes, and (2) locally rigid/planar parts of a non-rigid/non-planar instance. For example, among different views of the "plane flying" and "Brooklyn bridge tower" in Figure 1, relative positions of feature points stays the same for local near planar surfaces as well as for non-severe viewpoint changes; the butterfly in Figure 2 has non-rigid motion, but most of the local rigid sub-structures (e.g., the wing) still keeps their spatial layout consistent. We apply an elastic spatial consistency checking strategy to be able to accumulate evidence from these locally consistent patches in 3D view changes and non-rigid transformations.

Our model should be neither too weak to reject inconsistent spatial layouts nor too strong to rule out true spatial

[1] http://www.google.com/mobile/goggles/
[2] http://snaptell.com/

configurations. Specifically, the model should be able to (1) accumulate evidence from locally consistent patches and tolerant small motions/viewpoint changes for non-planar/non-rigid instances; (2) work reasonably for NDs; and (3) effectively filter inconsistent spatial configurations. Instead of modeling the transformation for absolute spatial locations, we use a "sketch-match" process to model the topology of spatial layouts of matched feature points. Since our method is based on Delaunay Triangulation, we name our spatial checking as DT in short.

Sketch: For instance search, given the matched words between a query instance $\mathcal{Q}$ and a reference image $\mathcal{R}$, DT sketches the spatial structures of $\mathcal{Q}$ and $\mathcal{R}$ respectively based on the matching locations. Figure 2 shows an example of the triangulation constructed on matched points of $\mathcal{Q}$ and $\mathcal{R}$.

For DT, it is a deterministic algorithm and the resulting triangles tend to be "regular", such that spatially neighbored points are coupled as edges and triangles, which are stable against small spatial perturbations as long as the topology holds. Note that the topology information is sketched into the graph after triangulation. For example, each edge (triangle) depicts the spatial nearness of two (three) points, and the full set of edges (triangles) gives a rather detailed "sketch" for the relative positioning of matched features. In this way, the absolute locations of the matched features are discarded and only the topology remains. Note this representation is invariant to scale, rotation changes.

For constructing meshes, the one-to-one mapping constraint needs to be enforced to ensure the number of nodes in each graph is identical. This is done by enforcing a point from $\mathcal{Q}$ to match only one point on $\mathcal{R}$ with the smallest Hamming distance. The enforcement effectively prevents an excessive number of redundant matches, an effect known as the "burstiness" [9].

Match: After triangulation, the spatial consistency is measured by graph matching. We have compared several strategies by considering different local structures (such as edges, triangles), and different weighting functions in terms of performance and efficiency, resulting in a computational efficient approach for matching graphs. With $\Delta\mathcal{Q}$ denoting the mesh of $\mathcal{Q}$, the geometric consistency of $\mathcal{R}$ and $\mathcal{Q}$ is measured as:

$$\mathrm{BF}(\mathcal{Q}, \mathcal{R}) = \|\mathbf{E}_{\Delta\mathcal{Q}} \cap \mathbf{E}_{\Delta\mathcal{R}}\|, \tag{1}$$

where $\mathbf{E}_{\Delta\mathcal{Q}}$ denotes the edge set of $\Delta\mathcal{Q}$, and BF indicates the number of common edges[3] between $\mathcal{Q}$ and $\mathcal{R}$. The retrieval score of $\mathcal{R}$ is then weighted by $\mathrm{BF}(\mathcal{Q}, \mathcal{R})$. This measurement works well in practice, because the features are coupled together while matching, resulting much lower false positive rate.

3.3 Discussion

After triangulation, the mesh can be regarded as an "abstract", or the approximation for the original shape. In computer graphics, this mesh is usually used to approximate the original shape. While in our case, the edge is encoded with the information of spatial topology for matched feature points. By viewing this mesh as a graph, the process of *sketch* discards the absolute spatial locations and leaves only the relative spatial nearness of a set of points distributed on a plane. Then the *match* process measures the topological layout consistency as graph similarity. Figure 2 gives an example on how DT works. Due to the non-rigid motion of the flipping wings, there are no linear transformations that could transform the matching locations from one to the other. If RANSAC is used with a linear model (either homography or epipolar), only a fraction of the "good" matches could survive, because the dominant linear transformation (e.g., defined by the matches in the yellow solid ellipse) will rule out many other good matches (e.g., matches on the wing in red dashed ellipse). For example, E-WGC is only able to locate five true matches for similarity ranking. DT, on the contrary, can accumulate evidences from both wings (yellow and red ellipses) and obtain a much higher confidence in topological similarity, since only relative positioning is sketched. Besides the locally consistent patches, non-rigid or non-planar regions of an object can also be partially tolerated as long as the motion of 3D structure is not severe. Interestingly, this assumption often holds for real life objects in practice. For example, when people are walking, spatial locations of each body parts only move in a small range, which would result in a graph different from that for irrelevant objects. In Figure 2, the high similarity score is also partially contributed by the this kind of topological consistency between wings.

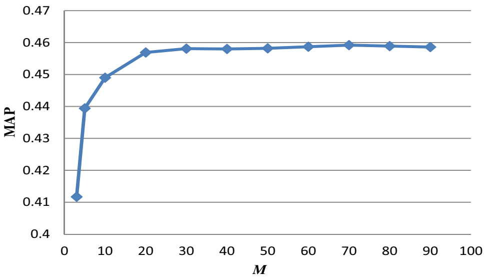

Figure 3: Sensitivity test of M on TV11, by applying DT on different number of max matching points.

While simple, DT has the following merits: (1) the relative spatial position of words is considered, (2) no assumption of any transformation model is made, (3) a certain degree of freedom for variations of word positions is allowed. Compared to weak spatial checking techniques, criterion (1) considers the topology of words, and thereby is more effective in measuring geometric consistency. Compared with strict spatial checking [23, 22], criterion (2) does not impose any prior knowledge on types of instances and transformations, and thus the checking of geometric coherency is looser. However, by allowing variations of local changes as stated by criterion (3), DT is a flexible model, which is more adaptable to INS. A fundamental difference between DT and other spatial checking techniques is that no pruning of false matches or model estimation is involved. Instead, DT enumerates the potential true matches with the local topology consistency based on criteria (1) and (3), while tolerating good matches by not imposing any prior constraints based on criterion (2).

3.4 Complexity

Time: The two major steps of DT are the triangulation and the counting of common edges. The first step can be efficiently conducted by divide-and-conquer in $O(n \log n)$ time, where n is the number of matched words between $\mathcal{Q}$ and $\mathcal{R}$. The second step can be done by a simple linear scan of

[3] Two edges are regarded as common if their vertices share the same visual words.

edges with $O(|e|)$, where $|e| = O(n)$ is the number of edges. So the computation is dominated by $O(n \log n)$. In our experiments, since a large vocabulary is used, n is quite small a number in most cases. Whenever n is larger than some value M, random sampling is performed to limit maximal M matching points, such that only a small random subset of matches is evaluated by Eq. (1). Figure 3 shows the sensitivity test on TV11 dataset. As shown, larger M gives more detailed sketch and better performance. When M is large enough, the performance tends to be stable. In our experiments, we set $M = 30$ to balance the efficiency and performance. In practice, DT runs fast, since it is only applied on images that have common visual words with the query image.

Space: For DT, we need to keep track of the matched points locations $[(q_x, q_y), (r_x, r_y)]$ between the query q and each reference image r. For a dataset with N images, $4 \times M \times N$ short integers are needed, which is approximate 288 MB, if $M = 30$ and $N = 10^6$ for a million scale dataset.

4. CONTEXT MODELING WITH “STARE"

Another problem for INS is how to use the background context. The ROI region alone gives clean and precise description for the target but less information, while the whole image carries more cues with more noises. The region inside ROI is definitely important, since it indicates the searching focus. However, we know little about the relevancy between the instance and the background context. Whether to use context information is by no means easy to tell, without the knowledge of the reference dataset beforehand.

Inspired by the perception behavior of human eyes, we weight features in different regions with a “stare” model to simulate human eye-sighting. At the time of starring at something, human eyes always have a *focus* (f), where things can be captured clear and nice, and scenes away from the *focus* “blurs” accordingly. In our “stare” model, the *focus* is a virtual point defined as the center of ROI, and the surrounding regions are down-weighted by a Gaussian function. The complete weighting function $k(x)$ for a feature x is given as:

$$k(x) = \begin{cases} 1, & \text{if } x \in \mathbf{ROI}, \\ \exp(-\frac{\|x-f\|^2}{2\delta^2}), & \text{otherwise}, \end{cases} \text{ with } \delta^2 = -\frac{diag^2}{8 \ln 0.1}, \tag{2}$$

where *diag* is the length of diagonal axis of the query image. Figure 4 (left) is an illustration with a circular ROI located at the center of a square picture. With the assumption of uniformly distributed feature points on the image plane, integrating the weights on ROI and background according to Eq. 2 gives the contribution ratio $f(r, N)$ of features inside the ROI while retrieving:

$$f(r, N) = \frac{\int_{x \in ROI} 1 \mathrm{d}x}{\int_{x \in \overline{ROI}} \exp(-\frac{\|x-f\|^2}{2\delta^2}) \,\mathrm{d}x + \int_{x \in ROI} 1 \mathrm{d}x} \tag{3}$$

Figure 4 (right) plots the simulation result of this ratio of contribution $f(r, N)$ with respect to the ratio of sizes ($2r/diag$). With the “stare” model, we adjust the weights on object and context adaptively for different sizes of instances. We tend to lay more emphasis on context for smaller instances, and vice verse. Note the curve in Figure 4 (right) is not completely sigmoid-like, since part of the ROI is out of the image when the ratio $\frac{2r}{diag} \in (\frac{\sqrt{2}}{2}, 1]$.

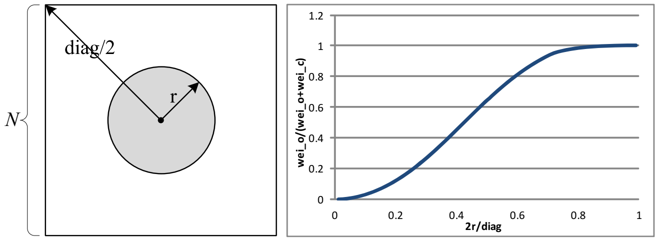

Figure 4: Left: illustration of the “stare” model with a circular ROI (with radius r) on a square image (resolution: $N \times N$). Right: the proportion of accumulated weights from ROI (the object), with respect to the ratio of sizes “$2r/diag$”.

Table 1: Statistics for TV11 & TV12.

	# topic	# query	# video	# frame	# feature
TV11	25	95	21k	574k	484m
TV12	21	102	75k	822k	1,323m

5. EXPERIMENT

This section evaluates the proposed methods for spatial checking (DT) and context modeling (“stare”). We start by introducing the TRECVID INS datasets, and then evaluate the performance of each method by comparing with other state-of-the-art techniques.

5.1 Dataset and Retrieval Model

Dataset: We use the TRECVID [17] INS datasets in years 2011 and 2012, named as TV11 and TV12, for experiments. The datasets contain video clips/shots cut from BBC Rushes and Flickr videos respectively for TV11 and TV12. The queries are topics on person, object, and location entities, which are delimited with several image examples together with the masks indicating the instances. The task [17] is to locate for each query topic up to the 1000 clips most likely to contain a recognizable instance of the entity. Figure 1 shows some query topics together with their corresponding ground truth video frames in the datasets. Table 1 summarizes the statistics of the datasets, and Table 2 lists the query topics in TRECVID datasets. Although there are only 25 (21) topics in TV11 (TV12), they cover a wide range of real life instances including objects, locations, and person. On average, each topic has 3.8 (4.9) image examples for TV11 (TV12), and a binary mask is also provided for each image example. Mean inferred AP (denoted as MAP for short) is used for evaluation.

Retrieval Model: Unless otherwise mentioned, the following BoW-based retrieval model is adopted for all the experiments. For offline processing, keyframes are extracted at the rate of one frame per second from raw videos, and Hessian-affine detector [12] and SIFT [11] descriptor are used for feature extraction. A hierarchical vocabulary [13] with 250k leaf nodes is constructed using K-Means in a top-down manner. Then, the features are indexed with an inverted file for fast retrieval. Auxiliary information, including Hamming signature [8] and spatial locations, are also indexed for word filtering and geometric checking. During online retrieval, a similar procedure is carried out for each query example. To reduce quantization error, a descriptor is assigned to multiple visual words by soft-weighting [15]. By

Table 2: Topic lists for TV11 & TV12 datasets.

TV11		TV12	
ID	Topic Name	ID	Topic Name
9023	setting sun	9048	Mercedes star
9024	upstairs, inside windmill	9049	Brooklyn bridge tower
9025	fork	9050	Eiffel tower
9026	trailer	9051	Golden Gate Bridge
9027	SUV	9052	London Underground logo
9028	plane flying	9053	Coca-cola logo - letters
9029	downstairs, inside windmill	9054	Stonehenge
9030	yellow dome with clock	9055	Sears/Willis Tower
9031	the Parthenon	9056	Pantheon interior
9032	spiral staircase	9057	Leshan Giant Buddha
9033	newsprint balloon	9058	US Capitol exterior
9034	tall, cylindrical building	9059	baldachin-St.Peter's Basilica
9035	tortoise	9060	Stephen Colbert
9036	all yellow balloon	9061	Pepsi logo - circle
9037	windmill seen from outside	9062	One WTO building
9038	female presenter X	9063	Prague Castle
9039	Carol Smilie	9064	Empire State Building
9040	Linda Robson	9065	Hagia Sophia interior
9041	monkey	9066	Hoover Dam exterior
9042	male presenter Y	9067	MacDonald's arches
9043	Tony Clark's wife	9068	PUMA logo animal
9044	American flag		
9045	lantern		
9046	grey-haired lady		
9047	airplane-shaped balloon		

traversing the index with HE filtering, images sharing common visual words are rapidly retrieved from the reference dataset. Since the final relevancy is evaluated at video level, the score for each video clip is obtained by accumulating scores from its keyframes, and the evidence of each query example is linearly combined by average fusion for the final ranking list.

5.2 Performance Comparison

We compare the following approaches: WGC (Weak Geometric Consistency) [8], E-WGC (Enhanced WGC) [23], GVP (Geometric-preserving Visual Phrases) [22], and our proposed approach DT. All the approaches are built on top of the BoW model described in Section 5.1. GVP is a voting approach that uses offset (or translation) information for rapid geometric checking. WGC, in contrast, votes the dominant scale and orientation for fast but weak geometric checking. E-WGC incorporates the advantages of GVP and WGC by voting the translation after scale and orientation compensation. We also test variants of DT by applying it on the whole query image (denoted as DT), instance only (denoted as DT_O), and context modeling using "stare" (denoted as DT_C).

5.2.1 Spatial Checking

Generally, in the set of matched points between two images using BoW method, there are many mismatches and outliers due to photometric and geometric variations. Applying a geometric consistency checking is important to remove false positives and improve the performance. Figure 5 contrasts the performances of different spatial checking techniques, and Figure 6 shows several search examples with the ranking information attached on the right side. The baseline method of BoW does not use any spatial checking techniques and ranks results purely based on visual similarity. WGC, E-WGC, GVP and DT, which impose a spatial consistency constraint on the matching points, show similar or better performances as BoW. In particular, WGC filters false matches by voting the dominant scale and orientation between two images, but makes no guarantee on consistent spatial layouts for the matching points. E-WGC considers the scale, orientation, and translation jointly into an affine model, and votes the dominant translation offsets after compensating the difference of scale and orientation.

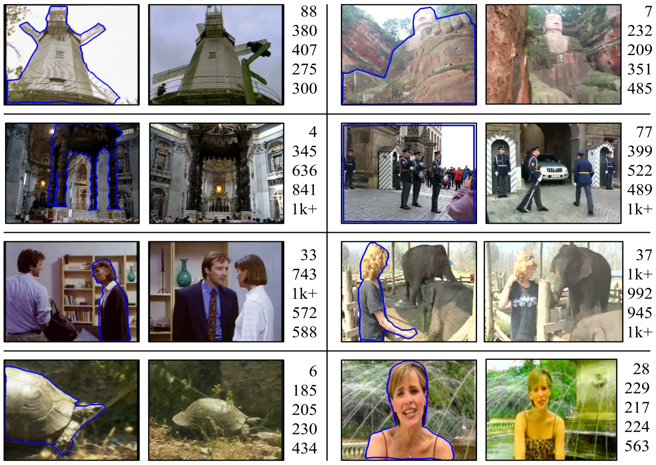

Figure 6: Examples on the ranks of retrieved images. For each example, the query is shown on left, and the corresponding retrieved images are on the right. The ranks of the retrieved image given by different spatial verification techniques are indicated by the numbers on the right hand side, ordered by DT, BoW, WGC, E-WGC, and GVP from top to bottom.

It works well for ND search, since scaling, rotating, and translating the image are common operations to generate NDs. However, both WGC and E-WGC suffer from imprecise scale/orientation estimation during feature extraction, especially for images with heavy noises, non-rigid objects, or 3D scenes captured from different viewpoints. GVP can be regarded as a special case of E-WGC, when two images are with identical scale and orientation. In other words, GVP votes the translation without compensating scale and orientation. Thus, it gets rid of the potential variations in scale/orientation estimation by assuming features from two images share the same scale and orientation, but also becomes more sensitive to scale/orientation changes even for ND pairs. Note that the spatial consistency model used in WGC, E-WGC, and GVP are all rooted in planar homography. Although they are able to rank some true responses (mainly near duplicates) higher, the final performance is downgraded because of the large number of falsely pruned true matches. This observation coincides with that in [18], where only a few topics benefits from the homography model and others does not. In our case, eight (seven) topics in TV11 (TV12) are improved by imposing the homography-based techniques, while other topics show similar or worse performance. Note for topics that totally violate the homography, the stronger the model it uses, the worse the performance is. For example, the topics 9026 (trailer) and 9057 (Leshan Giant Buddha) are with 3D objects viewed from different viewpoints, so they suffer less on WGC, which is a weak constraint, than on E-WGC and GVP, which are strong point-to-point transformations.

DT, instead of using a homography-based transformation, models the topology layout of matching points into a graph. It has several benefits. First, it is born to be invariant to scale/orientation changes. Since only the connectivity of nodes matters for a graph, scaling and rotating the image result in exactly the same graph. For example, the query and reference images shown in the last row of Figure 6 give

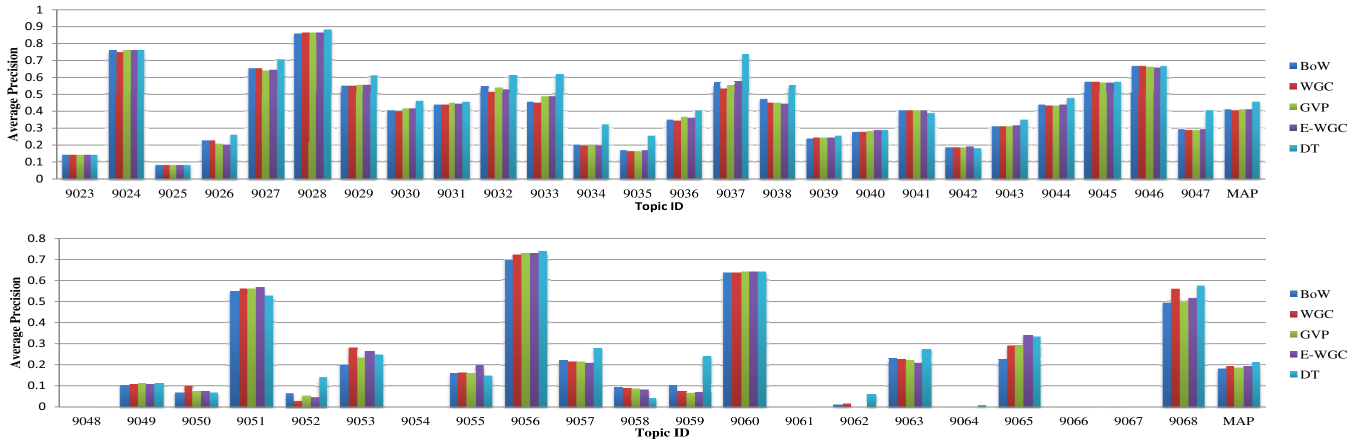

Figure 5: Performance comparison for different spatial checking techniques. Top: TV11. Bottom: TV12.

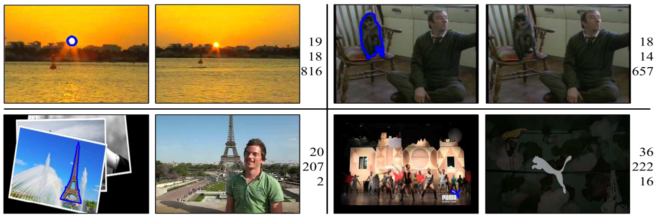

Figure 8: Effect of context modeling. The query and retrieved images are shown on the left and right sides, respectively. The ranking of different strategies on the retrieved images are indicated on the right hand side, ordered by the ranks given by applying context modeling with "stare" model (DT_C), whole image (DT), and instance only (DT_O).

the same graph, as long as the corresponding features can be matched. While for WGC and E-WGC, the requirement for small noises in the scale/orientation estimation makes them less robust in result ranking. Second, for non-homography spatial configurations introduced by different views of non-planar objects (first two rows of Figure 6) and non-rigid motions (3rd row), DT still get some evidence from the local topology-preserving regions. Third, for small number of the matching points caused by scale changes (left example in last row) or blur/noise/compression (right example in last row), DT actively boosts the ranking of the results according to Eq. 1, as long as the matched points are topologically consistent. While for other methods based on voting-and-pruning, the true responses with small number of matching point can only be boosted when the higher ranked false positives are downgraded by pruning of false positive matches. In other words, by (1) being invariant to scale/orientation changes, (2) allowing to get evidence from local topology-consistent sub-regions, and (3) act actively on boosting topology consistent results, true responses in INS have better chances to be boosted in the ranking list for DT than other homography-based methods.

5.2.2 *Context Modeling*

For TRECVID INS datasets, some instances in the reference videos appear in the same background context, while others in totally different context. Retrieving using either the whole image or the region inside ROI could miss some instances in the top list. Context modeling is designed to tradeoff between this situation and aims to bring both types of instances to top positions in the ranking list. Figure 7 shows the performance of different strategies of using context, including our context modeling with "stare", retrieving with the object inside ROI and whole image. We test these strategies on both methods without (BoW) and with (DT) spatial checking. By comparing the overall performances, the context does contain some useful information, since the method with BoW_O/DT_O gives worse result than BoW_C/DT_C on most topics. Our context modeling performs best among all variants for both BoW and DT, showing its effectiveness in retrieving instances with different background, without hurting too much on the ND results with the same background. Figure 8 further lists some examples, showing the tradeoff between exploring instances with different background and retrieving NDs with the same background. As shown in the 1st row, the ranks of ND results are only slightly downgraded by context modeling with "stare", since NDs usually cover the whole image and it still remains easy to retrieve with even down-weighted context information. While for instances appearing in different background as the query image (2nd row of Figure 8), the process of down-weighting of background context becomes essential to boost the rankings for results in different background context. The "stare" model shows the best performance for both BoW and DT, almost on every topics, except those with dense and discriminative features already, such as (2nd row of Figure 8) landmarks (9050: Eiffel tower, 9058: US Capitol exterior) and logos (9053: Coca-cola logo, 9068: PUMA logo). In such cases, adding more context confuses the targets that are already strong and clear. However, the ranks are not significantly downgraded and still remains top in the list, since the weighting in Eq. 2 still lay a lot of emphasis on instances with plenty of features.

5.3 Speed Efficiency

The experiments are conducted on a 8-core 2.67GHz computers with 30GB RAM. Only one core is used for online retrieval. Table 3 details the average running time for searching one query image from each dataset. As shown, BoW runs fastest among all the methods. WGC, GVP, and E-

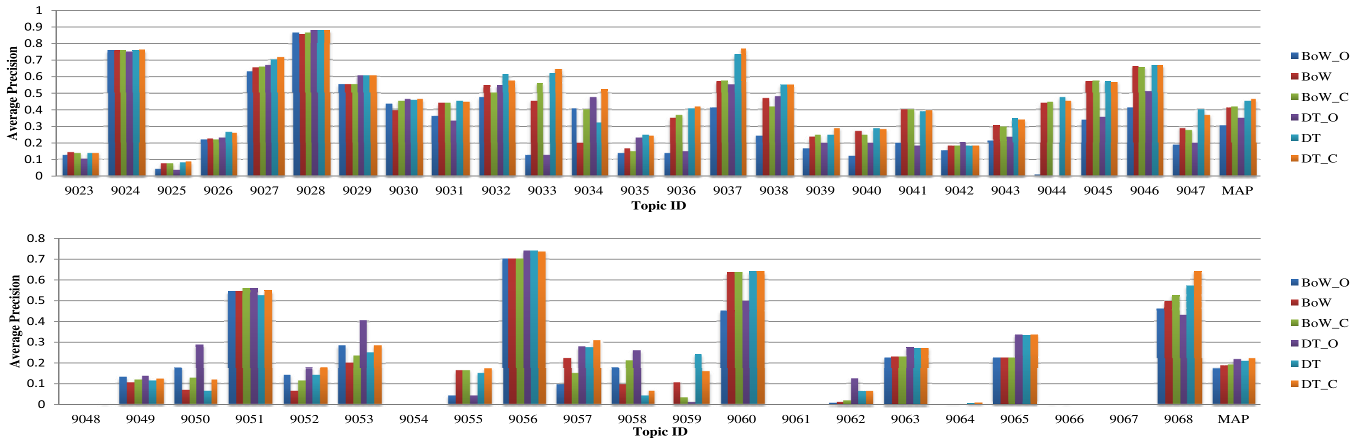

Figure 7: Performance comparison for different strategies of using context. Top: TV11. Bottom: TV12.

Table 3: The average running time (in milliseconds) for each method. The time includes feature quantization and online retrieval, but not feature extraction.

	BoW	WGC	GVP	E-WGC	DT	DT_C
TV11	173	232	227	292	217	219
TV12	625	885	851	989	875	880

WGC have a voting step to calculate the dominant transformation parameters, making them slower than BoW. DT and DT_C are also slower than BoW by introducing an extra step for triangulation and graph matching. Note the computation overhead for context modeling is negligible in practice. However, the extra time for DT is compensated by large performance gain. Note it takes much longer time for TV12 than TV11 in our experiments, since the number of features in TV12 is much more than that in TV11.

6. CONCLUSIONS

We have presented our approaches for searching instances from video collection, in the scenario of limited number of features for the query target and general spatial configurations. For INS, making better use of the limited information, including spatial and context cues, is critical for better performance. Specifically, DT, which improves the quality of visual matching by emphasizing the topology layouts of the matching points, boosts true results by accumulating evidence from local topology-preserving regions. To increase the amount of information for matching, context modeling via "stare" shows good tradeoff between exploration on instances with different background and exploitation on NDs with similar background. Our experimental result shows the effectiveness and efficiency of our methods for the problem of Instance Search.

7. ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118812).

8. REFERENCES

[1] http://en.wikipedia.org/wiki/Delaunay_triangulation.

[2] R. Arandjelovic and A. Zisserman. Efficient image retrieval for 3d structures. In *BMVC*, 2010.

[3] O. Chum and J. Matas. Large-scale discovery of spatially related images. *PAMI*, 32:371–377, 2010.

[4] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *CIVR*, 2007.

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In *ICCV*, pages 1–8, 2007.

[6] B. N. Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR*, (6):793–800, 1934.

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[8] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[9] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

[10] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[12] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60:63–86, October 2004.

[13] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[15] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[16] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[17] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, 2006.

[18] C. G. M. Snoek, K. van de Sande, A. Habibian, S. Kordumova, Z. Li, M. Mazloom, S. Pintea, R. Tao, D. Koelma, and A. W. M. Smeulders. The mediamill trecvid 2012 semantic video search engine. In *TRECVID*, 2012.

[19] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *ACMMM*, pages 218–227, 2007.

[20] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, pages 25–32, Aug. 2009.

[21] W. Zhang, L. Pang, and C. W. Ngo. Snap-and-ask: Answering multimodal question by naming visual instance. In *ACMMM*.

[22] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry preserving visual phrases. In *CVPR*, 2011.

[23] W. Zhao, X. Wu, and C. W. Ngo. On the annotation of web videos by efficient near-duplicate search. *TMM*, 2010.

[24] C. Zhu and S. Satoh. Large vocabulary quantization for searching instances from videos. In *ICMR*, 2012.