8-2015

Topological spatial verification for instance search

Wei ZHANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Topological Spatial Verification for Instance Search

Wei Zhang and Chong-Wah Ngo, *Member, IEEE*

***Abstract*—This paper proposes an elastic spatial verification method for Instance Search, particularly for dealing with non-planar and non-rigid queries exhibiting complex spatial transformations. Different from existing models that map keypoints between images based on a linear transformation (e.g., affine, homography), our model exploits the topological arrangement of keypoints to address the non-linear spatial transformations that are extremely common in real life situations. In particular, we propose a novel technique to elastically verify the topological spatial consistency with the triangulated graph through a "*sketch-and-match*" scheme. The spatial topology configuration, emphasizing relative positioning rather than absolute coordinates, is first sketched by a triangulated graph, whose edges essentially capture the topological layout of the corresponding keypoints. Next, the spatial consistency is efficiently estimated as the number of common edges between the triangulated graphs. Compared to the existing methods, our technique is much effective in modeling the complex spatial transformations of non-planar and non-rigid instances, while being compatible to instances with simple linear transformations. Moreover, our method is by nature more robust in spatial verification by considering the locations, rather than the local geometry of keypoints, which are sensitive to motions and viewpoint changes. We evaluate our method extensively on three years of TRECVID datasets, as well as our own dataset MQA, showing large improvement over other methods for the task of Instance Search.**

***Index Terms*—Instance Search, Spatial Verification, Non-planar and Non-rigid objects, Triangulated Graph.**

I. Introduction

Instance Search (INS) is a realistic problem initiated by TRECVID [1], which aims to retrieve any occurrences of the querying instance from a large video collection. The term *instance* here indicates a specific visual entity, e.g., a specific object, location, or person. Generally speaking, INS is featured by its definition of *instance level* relevancy. Different from traditional concept search where the relevancy is defined at semantic level, true responses of INS should depict the same instance. Different from similar image search which requires highly similar results, the search focus of INS is usually small and the relevant targets could exhibit different appearances as the query. Practically, INS is a fundamental problem for a wide range of applications, such as archive video search, law enforcement, personal video organization, browsing and brand-logo protection.

The challenge of Instance Search originates from the wide range of querying instances and complex capturing conditions. With reference to Fig 1, the difficulties can be summarized as: *small focus*, *viewpoint change*, *different background*, *non-planar* and *non-rigid targets*. For INS, the search target is usually small on both query and reference images. Accordingly, the same instance could show up in widely different background. Furthermore, the quality of SIFT [2] (or RootSIFT [3]) feature degenerates quickly as viewpoint changes. Finally, instance candidates to be retrieved could show non-rigid motions or non-planar surfaces, especially for 3D and non-rigid objects under different capturing conditions.

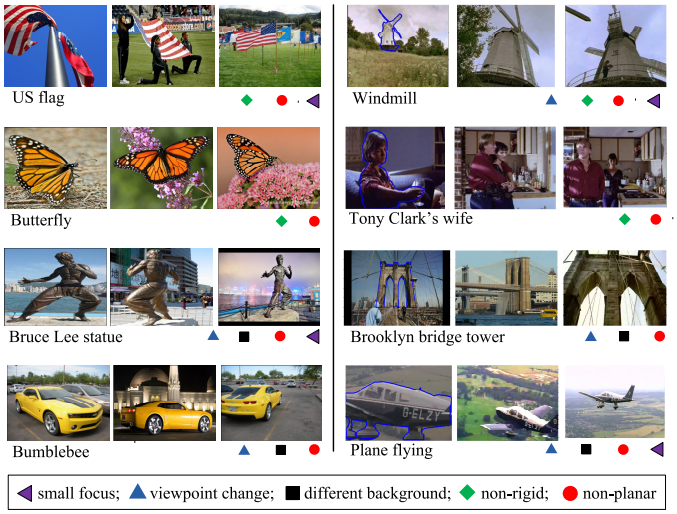


Fig. 1. Example instance groups in MQA (left) and TRECVID (right) dataset.

The state-of-the-art INS systems [4]–[6] are based on the BoW model [7], which was originally introduced for image search. Only a limited number of variants are introduced to amend BoW for addressing the aforementioned challenges in INS. These efforts include exploring the asymmetric nature of query and reference images [8], formulating query by weighting object and background [9], segmenting reference images into small object proposals [10]–[12] before indexing and searching.

Spatial verification in the context of INS is a vital component, since quality matching is highly demanded when the querying target is small and showing complicated spatial transformations. On one hand, much less information is available for small focus instances that only cover small image areas. On the other hand, local features are more unstable between relevant image pairs, especially for small non-planar and non-rigid instances viewed from different angles. Although previous studies [7], [13]–[16] have progressed a lot in spatial verification for similar image search, this component is still missing in the case of Instance Search. This paper studies the suitable spatial verification technique for Instance Search.

For traditional spatial verification methods, the common

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118812). We thank BBC for providing the EastEnders dataset: Programme material © BBC.

The authors are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: {wzhang35, cscwngo}@cityu.edu.hk).

practice is to assume a linear transformation between relevant image pairs. Such assumption works well for similar image search (e.g., copy detection, landmark search), where rigid structures and near-planar surfaces are abundance. However, such linear assumption does not hold for the queries in INS, e.g., non-rigid objects, multiple views of rigid objects, or non-planar scenes, as shown in Fig. 1. A better model should deal with the complex spatial transformations in INS. Here in this paper, *complex spatial transformation* refers to non-linear coordinate mapping between instances of the same object. Because of non-rigid motion or non-planar surface, the spatial relationship of the corresponding points goes beyond the simple linear transformation. For INS, previous models are either too strong [14]–[16] to tolerate the complex spatial transformation of relevant instances, or too weak [7], [13], [17] to reject random hits.

This paper proposes a triangulated-graph based spatial verification technique to emphasize the topological (i.e., relative ordering of 2D points) rather than a strict linear mapping between corresponding points. The technique improves instance matching by accumulating evidence from local topology-preserving patches for instances with complex spatial configurations, aiming for boosting the ranks of topological consistent results. In particular, we target to take better use of the limited information from small instances, by modeling the spatial configuration properly. In other words, the lack of information is compensated by quality matching via topology verification. Different from previous methods that impose a linear transformation over the absolute matching locations, we sketch and match the spatial topology based on a triangulated graph. In short, the focus of this paper is to explore spatial topology that ideally could lead to better accuracy for visual instance search without sacrificing much in speed efficiency. Spatial verification is imposed during search time, rather than as a post-processing step for fine tuning the search results. In other words, the topological consistency contributes directly to the scoring of instance similarity and is used for ranking of search results.

The main contribution of this paper is the proposal of an elastic spatial verification in the context of Instance Search. Triangulated graph is introduced to model the complex spatial transformations between relevant instances, which suits better for general Instance Search, where lots of non-planar and non-rigid instances are expected. At the same time, our method is also compatible with the spatial configuration for traditional similar image search, which makes our method suitable for general Instance Search. This manuscript extends upon our previous conference versions [9] and [18]. In this paper, we further improve our method by introducing a weighting strategy for more robust spatial verification. Moreover, more analysis on time complexity and noises in local geometry of SIFT, and more experiments on various datasets are included to better justify our method.

The remaining parts of this paper are organized as follows. Section II reviews related works on spatial consistency verification. Section III presents our technique on topological spatial verification. Section IV presents our experiments, and finally Section V concludes this paper.

II. Related Works

Since this paper focuses on spatial verification, we review existing BoW-based spatial constraints by categorizing them into *strong* and *weak* spatial consistency models. The original BoW model discards spatial information completely. Among all variants of BoW, spatial verification plays a critical role in addressing this limitation.

Strong models usually define a one-to-one point mapping from one image to the other. That is, these models are able to project the keypoints of one image to the corresponding locations on the other image. Most of the strong models are rooted in the homography geometry [19], which requires the scene under view to be planar or the camera centers fixed at a location. Philbin [14] reranks top retrieved results with three affine transformations, which are all special cases of the homography. Combined with RANSAC [20], these models work well for buildings with near-planar facades.

Zhao [15] extends WGC (Weak Geometric Consistency) [13] for spatial verification as E-WGC (Enhanced-WGC), where the points are back-projected using a simplified version of homography. Different from [14], this transformation is estimated using the local geometry (orientation and scale) of SIFT. GVP (Geometry-Preserving Visual Phrases) [16] is essentially the same as [15], except the orientation and scale are not considered when back-projecting the points. Avrithis [21] also adopts a similar spatial model as E-WGC [16], i.e., Hough voting in the transformation space. A different scoring method is used to weight multiple spatial transformations at different levels of spatial pyramid.

In general, strong models can be explicitly put as a linear transformation, where correspondences on two images can be related with a homography matrix (or its special cases). In practice, they work best on near-duplicate image retrieval [22], [23], where the spatial transformation can be modeled as a linear transformation.

Weak models does not assume a strict point-to-point mapping and thus has many diverse forms. In real life situations, most of the instances under query do not follow a strict linear transformation, and thus many weak models are studied to handle various instances.

For different views of rigid 3D objects, the essential spatial transformation is depicted by the epipolar geometry [19], where a fundamental matrix projects the points on one image to the epipolar lines passing through the corresponding points on the other image. Unlike homography, the fundamental matrix can only project a point to a line on the other image, which is a relatively weak constraint. Only a few works [24], [25] explore this model, which successfully retrieve some of the 3D structures.

Sivic [7] adopts a spatial model that favors clustered matched points, which only requires spatial closeness as a weak model. Similar in spirit, the approach in [10] pre-partitions images into thousands of object candidates to enforce the spatial closeness constraint. These methods are limited in spatial verification, since the spatial closeness is too weak to reject false positives. WGC [13] proposed by Jegou verifies the geometric coherency by voting the dominant scale

and orientation, and prunes outliers against the dominant transformations. Zhong [26] weakly checks the spatial consistency by verifying the horizontal and vertical ordering.

In general, weak models feature certain measurements for the distribution of matched points, e.g., density, mode, marginal distribution. In most cases, they are too loose to reject large number of false positives, especially in a large dataset.

As discussed before, spatial configuration in the context of INS is much more complex than traditional tasks, which is never addressed before. Existing methods are either too strong to tolerate the complex transformations, or too weak to reject random false matches that are common in large dataset. Non-planar (3D) structures violate the homography constraint, and non-rigid instances even do not comply with the epipolar geometry. Therefore we study a moderate model for the complex geometry in INS, which is loose to tolerate complex transformations in INS, while strong enough to reject false results.

III. Topological Spatial Verification

We first briefly introduce the spatial configurations involved in INS, and then present our method to address these situations. Finally, a detailed discussion is followed to analyze our triangulated-graph based technique.

A. Spatial Configuration

Let $\mathbf{Q}$ and $\mathbf{R}$ represent the homogeneous coordinates of feature points on the query $\mathcal{Q}$ and reference image $\mathcal{R}$, respectively. The corresponding spatial locations of two planar scenes can be related by a homography matrix $\mathbf{H}$, i.e., $\mathbf{Q} = \mathbf{HR}$. This formula defines a point-to-point mapping from different views of a plane. The commonly used affine transformation is a special case for homography, which works reasonably well for near-duplicate search (scaling, rotating, cropping) or near-planar instances (landmarks, paintings).

When it comes to 3D objects, the relationship between different views can be related by a fundamental matrix $\mathbf{F}$, i.e., $\mathbf{Q}^\mathbf{T}\mathbf{FR} = \mathbf{0}$. Unfortunately, the fundamental matrix defined by epipolar geometry can only map a point on the query to a straight line on the reference image, which is not sufficient for spatial verification. An even more complicated case is by considering non-rigid instances, for example the butterfly in Fig. 2. Previous analytical models are not applicable for being violated by the non-rigid motion.

To tackle these problems, we seek solutions from another perspective. Although there is no uniform transformation for the non-planar and non-rigid objects, the spatial topology tends to be stable for (1) different views of 3D objects, and (2) local rigid structures of a non-rigid object. For example, among different views of the 3D Bruce Lee statue in Fig. 1, relative positions of feature points stay the same for local near planar surfaces as well as for relatively small viewpoint changes. Similarly, although the butterfly in Fig. 2 shows non-rigid motion, some local rigid structures (e.g., each wing) still keep their spatial layout consistent. In the next subsection, we will propose an elastic spatial verification method, which is able to accumulate evidence from these locally consistent regions in 3D view changes and non-rigid transformations.

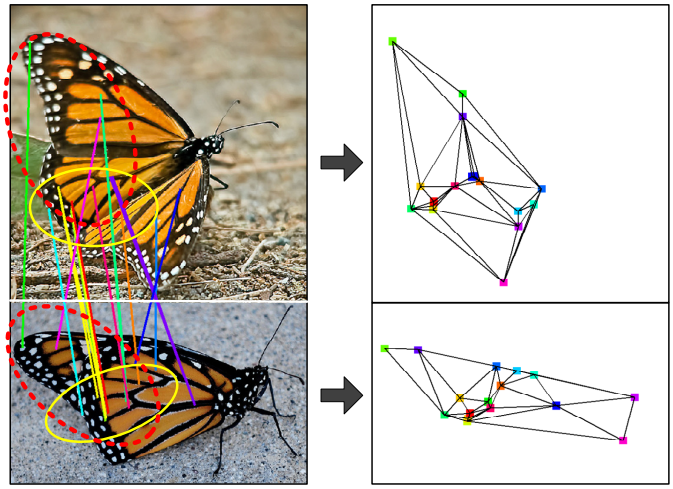


Fig. 2. The construction of the triangulated graphs based on matched visual words between two images. The matched visual words are indicated with the same color. Left: two images with their matched points lined up. Right: the triangulated graphs of the left images. This figure is best viewed in color.

B. Sketch-and-Match

Considering the complex spatial configuration as described in Sec. III-A, we need a model that is neither too weak to identify inconsistent spatial layouts nor too strong to rule out true spatial configurations. Specifically, the model should be able to (1) tolerant small motions and viewpoint changes, meanwhile accumulating evidence from locally consistent patches for non-planar and non-rigid instances; (2) work for highly similar images as well; and (3) effectively filter inconsistent spatial configurations.

The term *topology* is related to the properties of space that are preserved under continuous deformations including stretching, twisting and bending [27]. For example, a circle is topologically equivalent to an ellipse (into which it can be deformed by stretching). Topology can be used to abstract the inherent connectivity of objects while ignoring their detailed form. In this paper, we consider topology as a good property for modeling the complex spatial configurations in INS, and propose a "sketch-and-match" strategy for spatial topology consistency verification. Note that the 2D topology, also known as second-order topology, depicts the relative positioning of a 2D point set.

Sketch: We model the spatial topology using Delaunay Triangulation (DT) [28], [29]. DT is a technique used in computer graphics for building meshes out of a point set. DT couples points into triangles, in such a way that the edges between points are stable for small spatial variations as long as the nearby topology stays the same. For Instance Search, given all the matched words between the 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. Fig. 2 shows a real example of triangulation meshes constructed from matched visual words of $\mathcal{Q}$ and $\mathcal{R}$.

We apply the DT algorithm in [28] for sketching triangulated graph, by taking the set of matched points (or visual words) in an image as input. The triangulated graph is a planar graph where the addition of any edge will result in

a non-planar graph. During sketching, DT will avoid thin and skinny triangles as much as possible, by maximizing the sum of the minimum angles of all resultant triangles. Therefore, the topological layout is sketched as a triangulated graph, where the topology is approximated by the connectivity of points. For example, each edge (triangle) represents the spatial nearness of two (three) points, and the full set of edges (triangles) gives a "sketch" of the original topology of matching locations. In this way, the absolute locations of the matched words are discarded and only relative positioning is sketched. Note that this representation is insensitive to scale, rotation, and certain viewpoint changes.

To make the resultant graphs comparable, the one-to-one mapping constraint needs to be enforced. This is done by allowing a point from $\mathbf{Q}$ to match only one point on $\mathbf{R}$ with the smallest Hamming distance. Each point is indexed together with a 32-bit Hamming code [30] as feature for the distance measurement. The enforcement effectively prevents an excessive number of redundant matches, a problem known as the "burstiness" [31], which could corrupt the similarity scores when there are repeated patterns. Note that a large visual vocabulary is expected for minimizing the mismatches between points due to the enforcement of one-to-one mapping. The scheme, nevertheless, does not work when comparing images of repeated patterns or non-texture surfaces. The former will result in arbitrary matches of points, while the latter will end up with few or even no matching points.

Match: After triangulation, the spatial consistency is measured by graph matching. With $\Delta\mathcal{Q}$ denoting the triangulated graph of $\mathcal{Q}$, the geometric consistency of $\mathcal{Q}$ and $\mathcal{R}$, named *bonus factor*, 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 between $\mathcal{Q}$ and $\mathcal{R}$. Two edges are regarded as *common* if their vertices share the same visual words. The final score of $\mathcal{R}$ is then weighted by this factor. This measurement works well in practice, because the features are coupled together while matching, resulting in much lower false positive rate.

Besides the above graph matching strategy, we also compared an alternative choice, i.e., counting common triangles instead of edges (see Sec. IV-C for detail). However, we found that the strategy described in Eq. 1 works best for our case. Though simple, it gives efficient yet robust consistency estimation between two triangulated graphs. It is worth noting that our spatial verification is operated at the time of retrieval, not post-processing. Similar as WGC [30], we include the spatial locations (row and column) of each local point into the inverted file, so that spatial verification is performed along with searching.

C. Weighting Strategy

Furthermore, we introduce a weighting strategy to better estimate the spatial consistency. Essentially, each of the common edges should contribute differently to the bonus factor. Particularly, we measure the strength of each common edge E as the matching confidence of its both endpoints E_1 and E_2:

$$\mathrm{wBF}(\mathcal{Q},\mathcal{R}) = \sum_{E\in\{\mathbf{E}_{\Delta\mathcal{Q}}\cap\mathbf{E}_{\Delta\mathcal{R}}\}} \sum_{i=\{1,2\}} w(E_i, E_i'), \tag{2}$$

where $w(E_i, E_i')$ measures the matching confidence of the point E_i on the query image and its correspondence E_i' on the reference image. Recall that we index a 32-bit Hamming signature [30] for each local point. Thus $w(E_i, E_i')$ is defined as the binomial-based function:

$$w(E_i, E_i') = -\log_2\Big(\frac{1}{2^{32}} \sum_{j=0}^{d(E_i,E_i')} \binom{32}{j}\Big), \tag{3}$$

where $d(E_i, E_i')$ denotes the Hamming distance between the Hamming signatures of E_i and E_i'. This weighting function is motivated by the binomial distribution of Hamming distances ranging from 0 to 32. Highly similar matching points would lead to large weights, and thus the edges with high quality endpoints matches are emphasized accordingly.

D. Discussion

Anatomy of "sketch-and-match". For DT, the matched feature points on each image are first triangulated to approximate the spatial configuration with a graph. Then the consistency of topological layouts is measured by the similarity of the graphs accordingly. The process of *sketch* discards absolute spatial positions but keeps relative positioning of matching locations in the graph. Then the *match* process measures the topological layout consistency as graph similarity. Fig. 2 gives an example on how DT works. Due to the non-rigid motion of the flapping wings, there are no linear transformations that could transform the matching locations from one to the other. Techniques such as RANSAC [14], [20] can only keep a fraction of matches, in either of the wings that is locally rigid. In Fig. 2, the yellow ellipse encloses the six points retained by RANSAC, while the remaining seven valid matches are ruled out. In this 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 high similarity score of 0.67, since only relative positioning is sketched. Besides the locally consistent patches, non-rigid and non-planar parts of an object can be tolerated to certain degree as long as the motion is not severe. This assumption often holds for real life objects in practice. For example, the relative locations of each body part only move in a small range when a person walks. In Fig. 2, the high similarity (0.67) is also partially contributed by the topological consistency between wings.

Advantages of DT. While simple, DT has the following merits: (1) the relative spatial position of words is considered; (2) no assumption of any transformation model is enforced; (3) a certain degree of freedom for variations of matching positions is allowed. Compared to WGC [30], criterion (1) considers the topology of words, and thereby is more effective in measuring geometric consistency. Compared with strict spatial verification [15], 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) without the assumption of a transformation model, DT is a flexible model, which is more adaptable to non-rigid and non-planar instances under different capturing conditions. A fundamental difference between DT and other spatial verification methods 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). Since DT acts positively in finding true matches rather than negatively penalizing false matches, we name our measurement in Eq. (1) the "bonus factor".

Noisy local-geometry of SIFT. Some spatial verification methods (e.g., [15], [21], [30]) rely heavily on the local geometry (orientation, scale) of SIFT features. However, these methods are not stable based on our investigation of the noise and bias in SIFT's local geometry. On the contrary, our method only needs the location information of SIFT features, which is much precise and stable.

There are many noises in the scale and orientation estimation of SIFT. Wide baseline matching with SIFT is known to be difficult even for planar scenes, as the feature detector becomes vulnerable against large viewing angles [32]. Due to the non-rigid motions and non-planar surfaces that are common in INS, precisely estimating the local geometry is even challenging. As a result, recovering the transformation based on local geometry of SIFT becomes risky.

The local geometry of SIFT is also highly biased. Fig. 3 plots the statistics of the orientation (top) and scale (bottom) from the SIFT features extracted on 10k random Flickr images, using the Hessian Affine detector and SIFT descriptor. The most well-known implementation by Krystian Mikolajczyk [32], [33] is adopted. As shown, although the SIFT features are sampled from totally random images, the distributions of the local geometry are far from uniform-like, where strong biases are observed for both scale and orientation. Although the orientation bias can be partially explained as people's habit of photo capturing (i.e., favoring portrait and landscape shots), the severe bias in scale is mainly due to the limitation of geometric estimation. Therefore, estimating the spatial transformation with such bias is error-prone. In contrast, DT does not suffer from the noise and bias, since scale and orientation are not used for spatial verification.

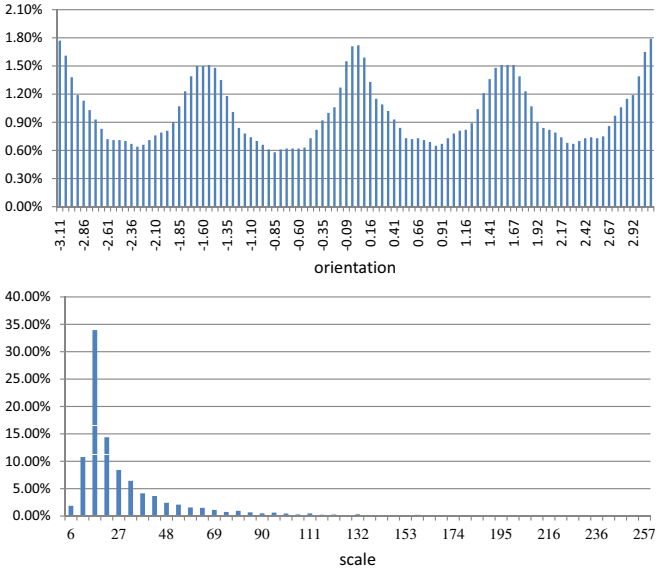


Fig. 3. Distributions of the orientation (top) and scale (bottom) of SIFT features extracted from 10k random Flickr photos.

E. Complexity Analysis

Time Complexity. 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 $\mathcal{O}(n \log n)$ time [34], where n is the number of nodes in the graph, i.e., the number of matched words between $\mathcal{Q}$ and $\mathcal{R}$. The second step can be done by a linear scan of edges with $\mathcal{O}(e)$, where e is the number of edges in the triangulated graph. Next, we show that e is $\mathcal{O}(n)$, and thus the total time complexity is dominated by $\mathcal{O}(n \log n)$.

According to Euler's formula, the following equation holds for any connected planar graph:

$$t - e + n = 1, \tag{4}$$

where t is the number of faces (triangles in our case) in a planar graph. On the other hand, since the graph is triangulated, we could derive another equation:

$$2e = 3t + k, \tag{5}$$

where $2e$ counts the number of oriented edges, $3t$ counts the number of edges associated to all triangles, and their difference is compensated by the number of edges (k) of the outer boundary[1]. It is easy to see from Eq. 4 and 5 that $e = 3n - k - 3$. Obviously, k is bounded by $\mathcal{O}(n)$, and thus $e = \mathcal{O}(n)$.

Overall, the computation is dominated by $\mathcal{O}(n \log n)$. In our experiments, since a large vocabulary is adopted, n is usually quite small. Whenever n is larger than some predefined threshold M, random sampling is performed to limit maximum M matching points, such that only a small random subset of matches is evaluated by Eq. (1). Here, M is defined as the maximum number of matched points for constructing a triangulated graph. Larger M sketches more details and gives better performance. When M is large enough, the performance tends to be stable. In our experiments, we set $M = 30$ to balance efficiency and performance. For small objects, the number of matched points is mostly less than 30, and this setting will not affect the matching of small objects in general. In practice, DT runs fast, since it is only applied on images that have common visual words with the query image.

Space Complexity. The space consumption is mainly for keeping track of the matched points locations $[(Q_x, Q_y), (R_x, R_y)]$ between the query $\mathcal{Q}$ and each reference image $\mathcal{R}$. For a dataset with N images, $4 \times M \times N$ short integers are needed, where M is the maximum sampled

[1]The number of edges for the outer ∞ face.

matched points for constructing the graph. Therefore, the space complexity is linear to N. Our method costs 288 MB memory when $M = 30$ and $N = 10^6$ for a million scale dataset. The space is negligible if compared to the memory consumed by inverted file ($\sim$15GB for $N = 10^6$ images) for indexing the visual words and other auxiliary information.

IV. Experiments

Experimental comparison was conducted against both the strong and weak spatial verification models, including the baseline BoW without any spatial verification, GVP (geometric preserving visual phrases) [16], WGC (weak geometric consistency) [30], and E-WGC (enhanced WGC) [15]. Note that the BoW baseline also includes components such as Hamming Embedding [30] and Multiple Assignments of words [35], which will be detailed in Section IV-B. We derive two versions of our approach: DT and DT*. DT is the original method using Eq. 1, while DT* is the extended version with Eq. 2. We name our approaches as DT for the use of Delaunay Triangulation for topology consistency checking. For fair comparison, all the tested approaches were implemented upon the same retrieval model described in Sec IV-B. The only difference is the use of the spatial model. BoW does not impose any spatial constraints, while GVP is a voting approach that uses offset (or translation) information for rapid geometric checking. WGC, in contrast, utilizes the dominant scale and orientation voting for fast but weak geometric verification. E-WGC incorporates the advantages of GVP and WGC by voting the translation after scale and orientation compensation.

We start by introducing the datasets (Sec. IV-A) and system framework (Sec. IV-B). Parameter settings are detailed in Sec. IV-C, followed by the performance comparison in Sec. IV-D and Sec. IV-E.

A. Dataset

MQA. A total number of 438 images are crawled from Flickr and Google Image, by querying 52 instances names (e.g., Wall Street Bull). This dataset[2] is originally used for visual instance naming [18], such that a wide range of real-life objects are covered, including fashion, vehicle, flower, pet, food, product, logo, landmark and art. Eventually, each visual instance has 5$\sim$15 (8.5 on average) image examples with different background (i.e., instance-level relevancy). For each of the instances, the first image is picked as the query, and the rest examples were then treated as the ground truth. In addition to the 438 images, we also construct a distracting dataset (Flickr1M) with one million images downloaded from Flickr by crawling "recent uploaded photos". No restrictions on tags, users, or locations are specified at the time of data crawling. Note that Flickr1M is not annotated, and is only used as a distracting set for scalability test. The left-hand side of Fig. 1 shows several examples of instance groups from the MQA dataset.

TRECVID (TV11$\sim$TV13). TRECVID [1] is an annual benchmark evaluation hosted by NIST for various video retrieval tasks. We use the INS datasets through years from 2011 to 2013 (denoted as TV11$\sim$13) for experiments. This dataset contains video clips cut from BBC Rushes (TV11), Flickr videos (TV12), and BBC EastEnders (TV13) as the reference set. The queries are usually objects, persons or locations provided by TRECVID, which are delimited with several image examples together with the masks indicating the instances. The INS task is to locate for each query topic up to the 1000 clips most likely to contain a recognizable instance of the query entity. Table I lists the query topics across three years, and Table II further details the data statistics. Note that the data size has increased significantly through years, and on TV13 we test our method on 4.5 million frames uniformly sampled at two frames per second from the reference videos. Some query images for TV13 can be found in Fig. 4.

[2]http://vireo.cs.cityu.edu.hk/mqa/

TABLE I
Topic lists for three years' TRECVID datasets.

TV11		9060	Stephen Colbert
ID	*Topic Name*	9061	Pepsi logo - circle
9023	setting sun	9062	One WTO building
9024	upstairs in windmill	9063	Prague Castle
9025	fork	9064	Empire State Building
9026	trailer	9065	Hagia Sophia inside
9027	SUV	9066	Hoover Dam outside
9028	plane flying	9067	McDonald's arches
9029	downstairs in windmill	9068	PUMA logo animal
9030	yellow clock dome	**TV13**	
9031	the Parthenon	9069	no-smoking logo
9032	spiral staircase	9070	red obelisk
9033	newsprint balloon	9071	Audi logo
9034	tall, cylindrical building	9072	Police logo
9035	tortoise	9073	cat face
9036	all yellow balloon	9074	cigarette
9037	outside windmill	9075	SKOE can
9038	female presenter X	9076	bust of Queen
9039	Carol Smilie	9077	this dog
9040	Linda Robson	9078	JENKINS logo
9041	monkey	9079	CD stand
9042	male presenter Y	9080	phone booth
9043	Tony Clark's wife	9081	black taxi
9044	American flag	9082	BMW logo
9045	lantern	9083	cafeteria
9046	grey-haired lady	9084	this man
9047	airplane-shaped balloon	9085	David magnet
TV12		9086	these scales
9048	Mercedes star	9087	VW logo
9049	Brooklyn bridge tower	9088	Tamwar
9050	Eiffel tower	9089	pendant
9051	Golden Gate Bridge	9090	wooden bench
9052	London subway logo	9091	Kathy's menu
9053	Coca-cola logo - letters	9092	this man
9054	Stonehenge	9093	turnstiles
9055	Sears/Willis Tower	9094	ketchup dispenser
9056	Pantheon interior	9095	trash can
9057	Leshan Giant Buddha	9096	Aunt Sal
9058	US Capitol exterior	9097	spheres
9059	baldachin-St.Peter's	9098	Parking sign

TABLE II
Dataset Statistics for TRECVID 2011$\sim$2013

dataset	# query	# ref clip	# ref image	data source
TV11	25	20,982	90K	BBC Rushes
TV12	21	76,751	822K	Flickr Video
TV13	30	469,539	4.5M	BBC EastEnders

Fig. 4. Query image examples in TV13 dataset. Instances under query are outlined with magenta contours. This figure is best viewed in color.

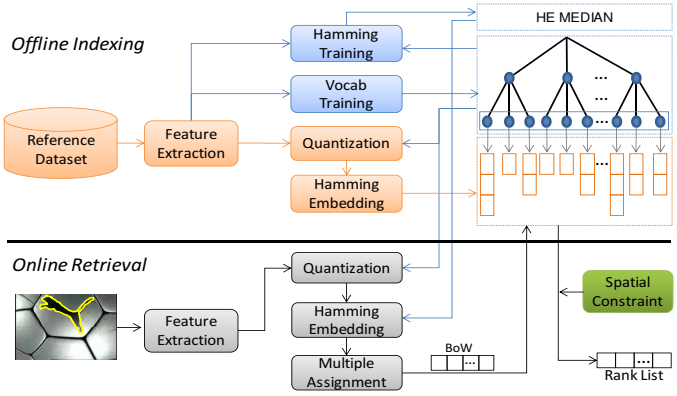


Fig. 5. Our retrieval model for Instance Search.

B. Retrieval Model

As shown in Fig. 5, our model is grounded on the recent advances in bag-of-visual-words representation (BoW) [7] and Hamming Embedding (HE) [30]. Initially a large vocabulary tree of one million leaf visual words is constructed via hierarchical K-Means. The implementation is based on [36], where local features (SIFT) are clustered hierarchically in a top-down manner, and a branching factor of 100 is used to split each non-leaf node. During offline indexing, SIFT features from the reference dataset are parsed down the tree until they reach the leaf nodes that best match the descriptors. Through this step, descriptors are quantized to their nearest visual words, and indexed into an inverted file for fast online retrieval. Auxiliary information, including the Hamming signatures and the spatial locations, are also indexed for early filtering and spatial verification, respectively. The Hamming signatures, represented as a binary vector of 32 bits, are generated by Hamming embedding [30]. During online retrieval, a similar procedure is carried out to process the query. To alleviate the adverse effect due to the quantization error, a descriptor is assigned to multiple visual words by soft-weighting [35], [37]. By traversing the index file with HE filtering, images sharing common visual words with the query are rapidly retrieved from the reference dataset. All experiments in this paper are based on this retrieval pipeline and the only difference is the way of spatial verification. For performance evaluation, we adopt the mean Average Precision (*mAP*) as the metric, where AP measures the area under the precision-recall curve.

In TRECVID dataset, each topic contains multiple query images. In our implementation, each query image example is searched against all keyframes from all clips. In other words, each query example is processed independently. To handle multiple keyframes from a clip, we max-pool the scores of keyframes from each video clip. Then the rank-lists from different query images are average-pooled as the final result. For evaluation, the final rank-list is truncated to a maximum of 1000 clips, following the same protocol as in TRECVID evaluation.

C. System Tuning

We first fine-tune our method on one of the dataset: TV11, and then freeze the settings for the rest of evaluation.

Triangulated Graph Matching. Here we test different choices for triangulated graph matching: (a) BoW (baseline method w/o any spatial verification); (b) EDGE (counting the number of common edges between two triangulated graphs); (c) TRI (counting the number of common triangles). BoW is included as a baseline. Both EDGE and TRI measure the graph similarity by accumulating evidence from local regularities of the graph, but they differ in the level of granularity. For EDGE, two edges are considered as common if both of their endpoints share the same visual words. While for TRI, the measurement is much strict by requiring one more pair of consistent endpoints as common triangles.

Table. III presents the mAP against different choices of graph matching methods. As observed, applying our topological spatial verification significantly improves the results, and EDGE shows clear advantage over TRI. This is because counting common triangles is usually too strict to tolerate small perturbations of instances with complex spatial transformations. On the other hand, EDGE is less sensitive to the noises by only requiring two pairs of consistent points. The following experiments will be based on EDGE, i.e., counting the number of common edges as the similarity of triangulated graphs.

Graph Sampling Threshold (M). Fig. 6 shows the sensitivity test of M, i.e., the maximum number of nodes sampled among the matched visual words for a triangulated graph, on TV11 dataset. In general, larger M implies denser sampling

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT TRIANGULATED GRAPH MATCHING STRATEGIES ON TV11.

	BoW	EDGE	TRI
mAP	0.4115	0.4586	0.4415

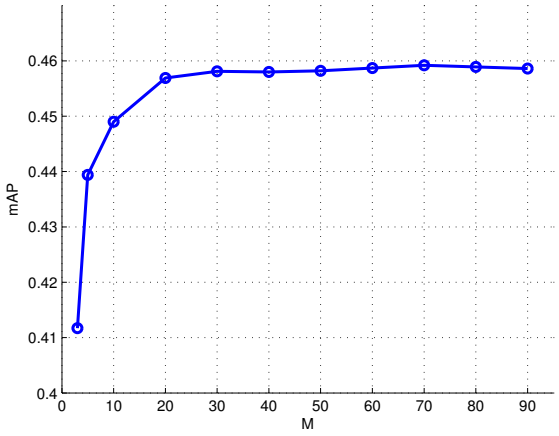


Fig. 6. Sensitivity test of M on TV11 dataset, by applying DT on different number of max matching points (M).

on the graph and sketches more details for the spatial configuration. Therefore the performance keeps increasing as M goes up. Interestingly, the mAP saturates around $M = 30$, and sketching more details does not give further improvement. For small objects, the number of matched points is often less than 30. For large objects on the other hand, using more points does not significantly boost the performance but slows down the speed. For near-duplicate images, the setting $M = 30$ still allows the construction of two highly similar (or identical) graphs for comparison. Thus we fix $M = 30$ in our following experiments to tradeoff between efficiency and accuracy.

D. Evaluation on MQA dataset

Fig. 7 shows the performance comparison on MQA dataset, by gradually adding more distracting images. Overall, DT and DT* consistently outperform previous methods across different scales, and more importantly, the margin gets larger as the scale approaches one million. This result indicates the robustness as well as the scalability of our approaches. We attribute this to the merits of triangulated graph in effective topology consistency measurement, resulting in better ranking of candidate instances with complex spatial transformations. Furthermore, by weighting the contribution from different edges, DT* gets even larger improvement. This is due to the effectiveness of wBF (Eq. 2) in identifying important edges.

Referring to Fig. 7, the performance is somewhat related to the strength of the spatial model. In MQA, relevant instances usually exhibit large variations. The performance usually gets worse when stronger model is adopted. BoW performs reasonably well on this dataset, and E-WGC gives worst result by enforcing a strict transformation. GVP, which relax the transformation by ignoring scale and orientation, generates similar mAP as BoW. On the other hand, weak models such

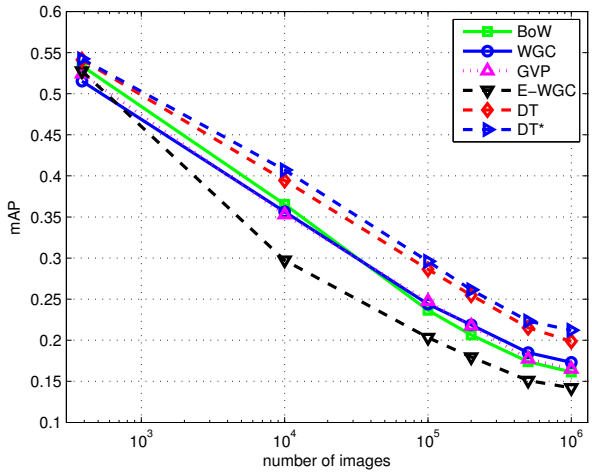


Fig. 7. Performance comparison of various approaches by adding more distracting images to the MQA dataset.

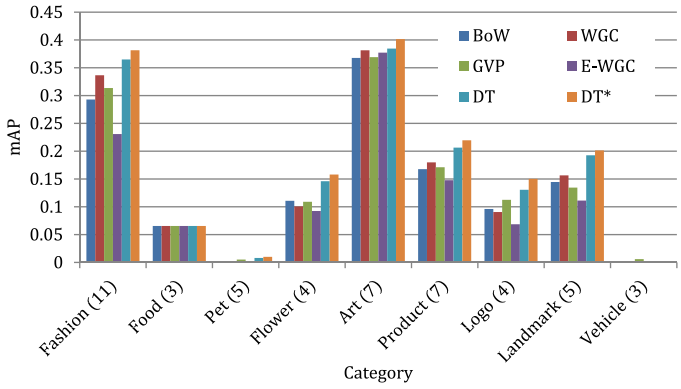


Fig. 8. The performance of different approaches on various instance categories. the number in parentheses indicates # queries for each category.

as WGC, actually improves the performance a little bit. DT and DT* give the best results by using a topology model with moderate strength.

By grouping the queries based on instance types as in Fig. 8, the performance gap is even more clear on non-planar and non-rigid instances such as Fashion, Flower and Product. However, Pet and Vehicle are still difficult to retrieve, because SIFT features are not stable for instances with fur or smooth (non-textured) surfaces.

Besides the non-planar and non-rigid example in Fig. 2, Fig. 9 gives more examples on how DT works for both relevant and non-relevant images. As shown, our method is also compatible to simple rigid and planar objects (left), since measuring the topology still makes sense in this case. It is worth noting that recovering the spatial transformation with RANSAC can be still difficult even for the left image pair, due to the mismatched points introduced by photometric and geometric variations. However, our method can still verify the consistency from image areas with valid matches. The example on the right shows the matching between the watch and a restaurant menu. WGC keeps as many as 21 matches, resulting in a misleading ranking. On the other hand, the irregularities

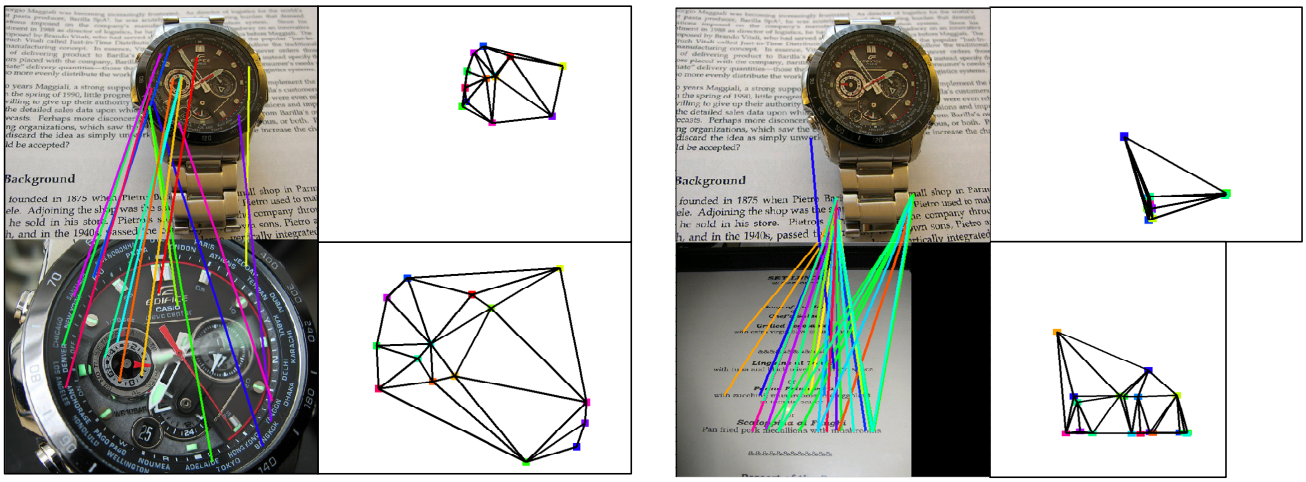


Fig. 9. More examples when matching with relevant (left) and non-relevant (right) images.

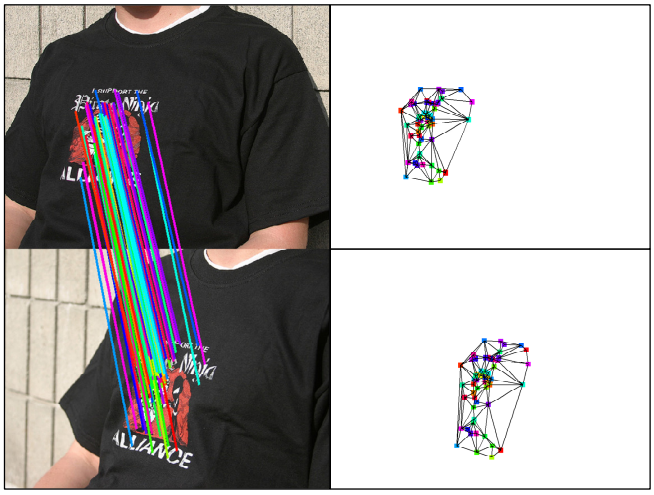


Fig. 10. An example when matching between near-duplicate instances.

in relative spatial positioning of matches result in two very different triangulated graphs. As a result, our methods rank the restaurant menu even lower in the result. Finally, our methods perform well despite using only a maximum of $M = 30$ matched points for constructing triangulated graphs. As shown in Fig. 10, randomly sampling 30 out of 85 points still gives two similar graphs.

E. Evaluation on TRECVID datasets: TV11∼TV13

Table IV summarizes the performance of different spatial verification methods on the TRECVID datasets. Fig. 11 further contrasts the detailed performance on each query topic, and Fig. 12 shows several search examples with the ranking information attached on the right side. The baseline method of BoW ranks results purely based on visual similarity. WGC, E-WGC, GVP, DT and DT*, which impose a spatial constraint on the matching points, show similar or better performances as BoW. Both WGC and E-WGC suffer from imprecise scale and orientation estimation during local feature extraction, especially for images with heavy noises, non-rigid objects, or

TABLE IV
PERFORMANCE (MAP) OF DIFFERENT SPATIAL VERIFICATION METHODS ON THE TRECVID DATASETS. NUMBERS IN PARENTHESES ARE THE MAP BY RUNNING RANSAC AS POST-PROCESSING OVER THE RESULTS OF BOW AND GVP.

	BoW	WGC	GVP	E-WGC	DT	DT*
TV11	0.411 (0.417)	0.407	0.411 (0.421)	0.411	0.459	**0.466**
TV12	0.185 (0.191)	0.195	0.189 (0.198)	0.195	0.212	**0.215**
TV13	0.143 (0.151)	0.143	0.148 (0.156)	0.151	0.198	**0.205**

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. Although it avoids the potential noises in the local geometry of SIFT features, GVP can not even handle near-duplicate images with scaling or rotation.

Although most approaches could rank true responses (mainly near-duplicates) higher, the final performance is downgraded because of the large number of falsely pruned relevant images. This observation coincides with that in [38], where only a few topics benefits from the homography model and others does not. In our case, only 8/7/11 topics in TV11/TV12/TV13 are improved by imposing the homography-based techniques, while other topics show similar or worse performance. Note that for topics that totally violate the homography, the stronger the model it uses, the worse the performance is. For example, the topics 9026 (trailer), 9059 (baldachin) and 9090 (wooden bench) consist of 3D objects viewed from different viewpoints. The problem is less severe for WGC (a weak constraint) than E-WGC and GVP (strong point-to-point transformations).

We include two additional runs based on RANSAC [14]. Specifically, RANSAC acts as a post-processing step for BoW and GVP, by reranking the top-200 returned results. In addition, we also adopt the "early stop" strategy [14] by terminating re-ranking if we process 20 images in a row without a successful verification. While RANSAC is capable of filtering false positives, it does not work well for 3D and non-rigid instances. As shown in Table IV, RANSAC

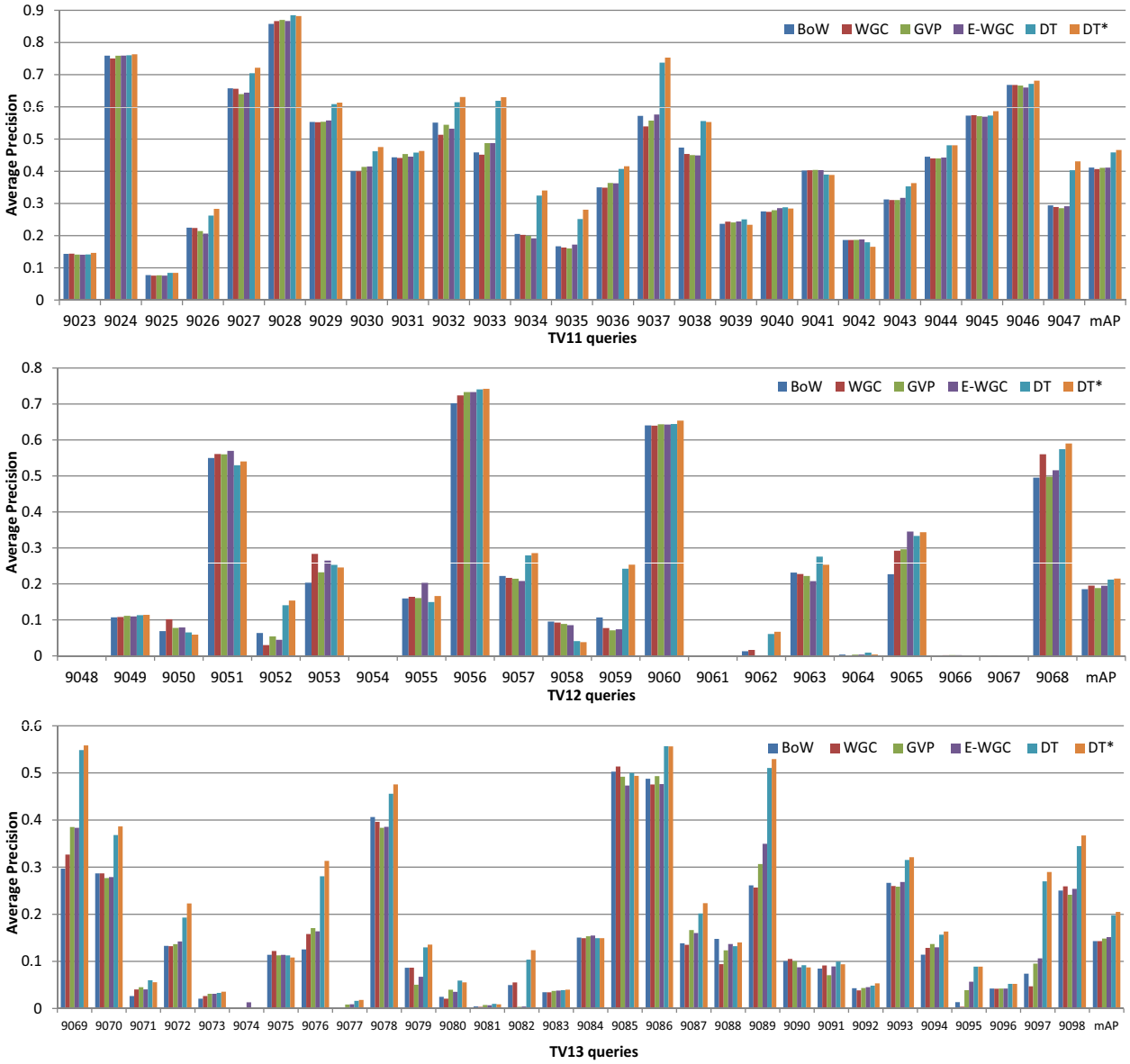


Fig. 11. Performance comparison for different spatial verification techniques. Top: TV11; Middle: TV12; Bottom: TV13.

only slightly improves the performances of BoW and GVP. Basically, RANSAC is effective in improving queries with near-duplicate instances, for example topics 9078 (a JENKINS logo) and 9085 (David magnet). However, when query and reference instances exhibit complex spatial configurations, such as topics 9076 (bust of queen) and 9086 (these scales), RANSAC can hardly reach a consensus with the single linear transformation. Only a small subset of features on a small planar or rigid part can be fitted by the estimated linear transformation.

DT and DT*, which models the topology layout of matching points into a graph, enjoy several benefits. First, our methods are born to be invariant to scale and orientation changes. Since only the connectivity of nodes matters for a graph, scaling and rotating an image result in exactly the same graph. For example, the query and reference images shown in the last row of Fig. 12 give the same graph, as long as the corresponding features can be matched correctly. While for WGC and E-WGC, the requirement for precise scale and orientation estimation makes them less robust in ranking. Second, for non-homography spatial configurations introduced by different views of non-planar objects (first two rows of Fig. 12) 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 the last row) or blur/noise/compression (right example in the last row), DT actively boosts the ranking of the results, as long as the matched points are topologically consistent. While for other methods based on voting-and-pruning, true responses with small number of matching points can only be boosted when the higher ranked false positives are downgraded by pruning of false positive matches. In brief, by (1) being invariant to scale and orientation changes, (2) allowing to get evidence from local topology-consistent sub-regions, and (3)

Fig. 12. Example ranks of retrieved images, including different views of non-planar instances (top two rows), non-rigid instances (3rd row), and instances with change of scale/blur/compression/noises (last row). Each example includes a pair of images, where the query is on left, and a retrieved image is shown 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*, DT, BoW, WGC, E-WGC, and GVP from top to bottom. For example, the numbers for the top-left example mean the retrieved image is ranked at the 78^{th}, 88^{th}, 380^{th}, 407^{th}, 275^{th}, and 300^{th} positions by DT*, DT, BoW, WGC, E-WGC, and GVP, respectively.

acting actively on boosting topology consistent results, true responses in INS have better chances to be upgraded in the ranking list for DT than other homography-based methods.

Fig. 13 gives more examples showing the strength and limitation of DT-based approaches, in comparison to WGC, E-WGC and GVP. DT shows clear advantage on non-planar instances such as topic 9057 (Leshan Giant Buddha) and non-rigid instances such as topic 9037 (windmill seen from outside), by tolerating complex spatial transformations. As other methods, DT performs equally well on topics with near-duplicate instances, such as topic 9031 (the Parthenon), where true positive instances only vary slightly due to small changes in lighting condition and noise. However, for instances showing large variations in visual appearances that result in excessive number of false word matches, DT performs equally poor as others. As shown in Fig. 13, topic 9054 (the Stonehenge) is one such typical example. Finally, DT suffers more on instances with repeated patterns due to the lack of unique one-to-one word matches. As a result, the resulting triangulated graphs could be considerably different. Weak models such as WGC deal better for these cases due to the use of histogram comparison rather than point-to-point matching. In the examples shown, i.e., topics 9058 (US Capitol exterior) and 9090 (wooden bench), WGC performs better because the orientation histograms of SIFT are still able to show similar distributions.

F. Speed Efficiency

The experiments were conducted on an 8-core 2.67GHz machine with 128GB memory. Only one core was used for online retrieval. Table V 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-WGC have a voting step to calculate the transformation parameters, making them slower. Compared to RANSAC, nevertheless, these models are still efficient. In our experiment, even when post-processing the top-200 results of BoW with "early stop" strategy [14], RANSAC still takes around 12 seconds on average for reranking. DT and DT* are also slower than BoW by introducing an extra step to construct and match the triangulated graphs. However, the extra time for DT and DT* is compensated by the large performance gain. Also note that it took much longer time on TV13 in our experiments, since the number of frames in TV13 is much more than that in other datasets (see Table II for details).

TABLE V
THE AVERAGE RUNNING TIME (IN SECONDS) FOR EACH METHOD. THE TIME INCLUDES FEATURE QUANTIZATION AND ONLINE RETRIEVAL, BUT NOT LOCAL FEATURE EXTRACTION. NOTE THAT THE TIME REPORTED ON MQA IS WITH ONE MILLION DISTRACTING IMAGES ADDED.

	BoW	WGC	GVP	E-WGC	DT	DT*
MQA	0.65	0.88	0.89	0.90	0.88	0.89
TV11	0.17	0.23	0.23	0.29	0.22	0.23
TV12	0.63	0.89	0.85	0.99	0.88	0.90
TV13	2.16	3.05	3.00	3.45	3.34	3.45

G. Discussion and comparison with other INS approaches

Over the years of TRECVID benchmark evaluations, there have been several branches of approaches experimented. These approaches are built upon different baselines and focus studies on different aspects of INS, and hence are not directly comparable with the work presented in this paper. Here we give a brief discussion of these approaches in comparison to our work.

BoW based model. Most of the successful approaches in INS are built upon the BoW model. For instance, the work in [6] experimented soft assignment of visual words and query expansion, which introduce 13% of improvement over the BoW model on TV13 dataset. Similar in spirit, a localized object search algorithm was proposed in [11] to rerank the initial results by the BoW model. Our work exhibits better retrieval performance than [6], [11] on TV13 dataset, and more importantly, introduces a larger degree of improvement (43%) over our BoW baseline.

Learning based retrieval. Peng [39] adopted multiple features and multi-bag SVM to train a model for each querying instance and classify the reference shots, which reports the mAP of 0.231 for TV12. This learning based approach is essentially effective in retrieving with only a few query examples, and thus generates similar or slightly better performance compared to our method. However, the learning-based approach is much slower compared to our method, since learning the model for each query instance is time-consuming and classifying each reference shot is essentially linear to the size of reference set.

Feature pooling and asymmetrical dissimilarity. The state-of-the-art performances were also reported in [4] and [40], which reach the mAP of 0.501 for TV11 and 0.313 for TV13. The success is based on average pooling of local features for

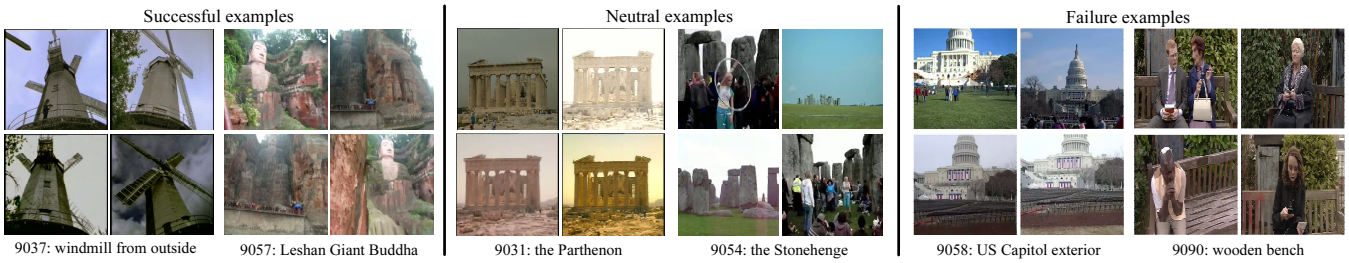


Fig. 13. Example visual instances that the DT-based approaches perform better (left), similar (middle) and worse than other spatial verification methods (WGC, E-WGC, GVP). Left: DT handles better for non-rigid motions (topic 9037) and non planar surfaces (9057). Middle: all methods perform well on near-duplicate instances (9031), but fail on instances with large appearance variations (9054). Right: DT suffers more on repeated patterns due to noisy word matching (9058 and 9090).

each shot, and the asymmetrical dissimilarity measurement between the query and reference shots. The pooled feature serves as a more robust representation, and the asymmetrical measurement penalizes more heavily on the reference images which do not have visual words matched against a given query. We used the best result provided by [4] as the input to our system, and further applied our topological spatial verification to rerank the initial search results. On TV13 dataset, our approach improves the mAP of [4] from 0.313 to 0.335, which is the best reported performance so far on this dataset to the best of our knowledge.

V. Conclusion

Spatial verification in the context of Instance Search is an important yet challenging problem. Both too strong and too weak constraints would fail on instances with complex spatial configurations. In this paper, we proposed a topological spatial verification method based on triangulated graphs, to explore the sugar spots between strong and weak constraints. For instances exhibiting complex spatial configurations, we explore from the perspective of topology, which is invariant to various spatial transformations. Our approach is featured by its topology modeling instead of traditional geometric mapping. On one hand, DT benefits from being insensitive to local spatial variations in BoW matches. On the other hand, it is sensitive to severe changes and repeated patterns that corrupt the topology layout. Our extensive experimental result shows the effectiveness and efficiency of our method for the problem of Instance Search.

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trevid," in *ACM Multimedia Information Retrieval*, 2006.

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

[3] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[4] D. Le, C. Zhu, S. Poullot, and S. Satoh, "National institute of informatics, Japan at TRECVID 2011," in *TRECVID*, 2011.

[5] W. Zhang, C.-C. Tan, S.-A. Zhu, T. Yao, L. Pang, and C.-W. Ngo, "Vireo @ trecvid 2012: Searching with topology, recounting will small concepts,learning with free examples," in *TRECVID*, 2012.

[6] Z. Zhang, R. Albatal, C. Gurrin, and A. F. Smeaton, "Trecvid 2013 experiments at Dublin City University," in *TRECVID*, 2013.

[7] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477.

[8] C.-Z. Zhu, H. Jégou, and S. Satoh, "Query-adaptive asymmetrical dissimilarities for visual object retrieval," in *International Conference on Computer Vision*, Oct. 2013.

[9] W. Zhang and C.-W. Ngo, "Searching visual instances with topology checking and context modeling," in *ACM International Conference on Multimedia Retrieval*, 2013.

[10] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Locality in generic instance search from one example," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[11] 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 at trecvid 2013: : Searching concepts, objects, instances and events in video," in *TRECVID*, 2013.

[12] Z. Li, E. Gavves, K. E. A. van de Sande, C. G. M. Snoek, and A. W. M. Smeulders, "Codemaps segment, classify and search objects locally," in *IEEE International Conference on Computer Vision*, 2013.

[13] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision*, 2008.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[15] W. Zhao, X. Wu, and C. W. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Transactions. on Multimedia*, vol. 12, no. 5, pp. 448–461, 2010.

[16] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry preserving visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[17] B. Fernando, E. Fromont, and T. Tuytelaars, "Mining mid-level features for image classification," *International Journal of Computer Vision*, vol. 108, no. 3, pp. 186–203, 2014.

[18] W. Zhang, L. Pang, and C.-W. Ngo, "Snap-and-ask: Answering multimodal question by naming visual instance," in *ACM Multimedia*, 2012.

[19] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] Y. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *International Journal of Computer Vision*, vol. 107, no. 1, pp. 1–19, 2014.

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

[23] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *International Conference on Image and Video Retrieval*, 2007.

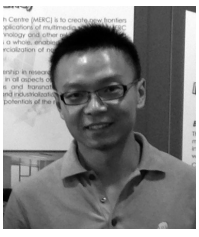
[24] R. Arandjelovic and A. Zisserman, "Efficient image retrieval for 3d structures," in *The British Machine Vision Conference*, 2010.

[25] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 371–377, 2010.

[26] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, Aug. 2009, pp. 25–32.

[27] http://mathworld.wolfram.com/Topology.html.

[28] B. N. Delaunay, "Sur la sphère vide," *Bulletin of Academy of Sciences of the USSR*, no. 6, pp. 793–800, 1934.
[29] Y. Kalantidis, L. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *ACM International Conference on Multimedia Retrieval*, April 2011.
[30] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 192–212, May 2010.
[31] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[32] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *Int. J. Comput. Vision*, vol. 65, no. 1-2, pp. 43–72, Nov. 2005.
[33] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[34] G. Leach, "Improving worst-case optimal delaunay triangulation algorithms," in *In 4th Canadian Conference on Computational Geometry*, 1992.
[35] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[36] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.
[37] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, pp. 42–53, 2010.
[38] 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.
[39] Y. Peng, X. Zhai, J. Zhang, L. Huang, N. Li, P. Tang, X. Huang, and Y. Zhao, "PKU_ICST at trecvid2013 : Instance search task," in *TRECVID*, 2013.
[40] D.-D. Le, C.-Z. Zhu, and S. Satoh, "National institute of informatics, japan at trecvid 2013," in *TRECVID Workshop*, 2013.



Wei Zhang received his B.Eng and M.Eng degrees from Tianjin University, Tianjin, China, in 2008 and 2010, respectively. He received the Ph.D degree from Department of Computer Science in City University of Hong Kong, Kowloon, Hong Kong, in 2015.

He was a Visiting Student in DVMM group of Columbia University, New York, NY, USA, in 2014. He was a former member of CV-lab in Tianjin University from 2008 to 2011. His research interests include computer vision, multimedia and digital forensic analysis.

Chong-Wah NGO received the M.Sc. and B.Sc. degrees in computer engineering from Nanyang Technological University of Singapore, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science & Technology, Clear Water Bay, Hong Kong.

He was previously a Postdoctoral Scholar with the Beckman Institute, University of Illinois in Urbana-Champaign, Champaign, IL, USA. He was also a Visiting Researcher at Microsoft Research Asia, Beijing, China. He is currently a Professor with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. His recent research interests include large-scale multimedia information retrieval, video computing, multimedia mining, and visualization.

Dr. Ngo was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2011-2014). He was Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and the Pacific Rim Conference on Multimedia 2014. He also served as Program Co-Chair of ACM Multimedia Modeling 2012 and ICMR 2012. He was the Chairman of the Hong Kong Chapter of ACM from 2008 to 2009.