NTT at TRECVID 2015: Instance Search

Xiaomeng Wu[*], *Taiga Yoshida*[**], *Jun Shimamura*[**], *Hidehisa Nagano*[*], *Kunio Kashino*[*], *Takahito Kawanishi*[*], *Kaoru Hiramatsu*[*], *Takayuki Kurozumi*[**], *and Tetsuya Kinebuchi*[**]

[*] *Communication Science Laboratories, NTT Corporation*
[**] *Media Intelligence Laboratories, NTT Corporation*

Abstract

This report describes our system created for the instance search task of TRECVID 2015. This system was tuned with the topics that were used for the instance search task of TRECVID 2014 and the BBC EastEnders dataset. The system was built on top of a standard local feature-based framework in which two spatial verification methods were newly incorporated. The first method called ensemble of weak geometric relations (EWGR) imposes a unified collection of pairwise geometric constraints on scale-invariant feature transform (SIFT) correspondences. The second method called Angle Free (AF) detects affine-SIFT correspondences and employs Hough voting in a 3D camera motion space. Specifically, the combinations of methods in our runs are as follows, where BOVW indicates the bag-of-visual-words model:

Run ID	**BOVW**	**EWGR**	**AF**
F_A_NTT_1	√	√	√
F_A_NTT_2	√	√	√
F_A_NTT_3	√		√
F_A_NTT_4	√	√	

Among the runs, F_A_NTT_1 and F_A_NTT_2 achieved the highest mean average precisions (MAPs), followed by F_A_NTT_3. F_A_NTT_4 underperformed the others in terms of MAP but was still superior to BOVW without spatial verification. From the results, we can see that both EWGR and AF contribute to the rejection of mismatches in local feature-based instance search, while AF is more effective against large 3D viewpoint changes. Overall, we found that EWGR was more suitable for discriminating and fast spatial verification and AF was better at handling large 3D viewpoint changes. The two methods were complementary and so their combination resulted in further improvement as regards the instance search task.

I. Introduction

The problem of instance search (INS) defined in TRECVID [10] is to rank database videos according to the probability of the existence of specific objects delimited by regions in a set of query images. In this report, we investigate the effectiveness of spatial verification for solving this problem. Although spatial verification has been widely proved to be successful in local feature-based image retrieval [2, 11, 13, 16], Zhu et al.'s study [18] is one of the first studies to report the contribution of spatial reranking to instance search from videos. Zhu et al.'s effort was admirable, however their method [18] based on random sample consensus (RANSAC) [11] is still founded on a compromise reranking framework because of its limited efficiency. Meanwhile, the authors did not consider the sensitivity of spatial verification in terms of large 3D viewpoint changes.

In this report, we attempt to determine whether or not the use of spatial verification supports a full search of a very-large-scale video database, namely one with 9.8 million keyframes extracted from 464 hours of

video. At the same time, we consider the severe problem of large 3D viewpoint changes: local features such as scale-invariant feature transform (SIFT) [6] are invariant to anisotropic transformations only to a limited extent and fail to match when a non-planar object is photographed under greatly differing camera angles. To address these issues, we newly incorporate two spatial verification methods in the framework of local feature-based instance search. The first method is called ensemble of weak geometric relations (EWGR) [16] and imposes a unified collection of pairwise geometric constraints on SIFT correspondences for discriminating and fast spatial verification. The second method, which is called Angle Free (AF) [14], detects affine-SIFT (ASIFT) [8] correspondences and employs Hough voting in a 3D camera motion space to handle 3D viewpoint changes.

The remainder of this report is organized as follows: Section II introduces the TRECVID INS benchmark and provides an overview of our system. Sections III, IV and V describe the three methods, namely the bag-of-visual-words (BOVW) model [11], EWGR and AF, whose combination we tested in our submitted runs. In the same sections, we provide details of our implementation and configuration, and report our results on the topics of both INS 2014 and INS 2015. We conclude the report in Section VI.

Note that the collection of ground-truth data provided by TRECVID constitute only a portion of master shots, i.e. there are many unlabeled master shots. For INS 2014, we removed all of the unlabeled master shots from the database, and then collected the top 1,000 results for the computation of the mean average precisions (MAPs). This is totally different from the official configuration in which unlabeled master shots are regarded as incorrect results. For INS 2015, we used the official configuration in our experiments.

II. Overview

A. TRECVID INS Benchmark

The INS benchmark is challenging because it was designed to simulate real demands for the multimedia retrieval community. The task description [10] is: given a collection of test videos, a master shot reference, and a collection of topics (queries) that delimit a person, object, or place entity in some example video, locate for each topic up to 1,000 shots that are most likely to contain a recognizable instance of the entity.

INS 2014 and INS 2015 use the same collection of test videos known as BBC EastEnders but two different collections of topics. There are 244 test videos with total duration of around 464 hours. According to the given master shot reference, we have 471,526 master shots, each of which lasts an average of three to four seconds. In our experiments, we sampled keyframes with a minimum frame rate of 6 frames per second and obtained 9,752,650 keyframes in total. Each INS task has 30 topics each consisting of four images. For each image there is a binary mask of the region of interest (ROI), as bounded by a single polygon. The topics are very challenging as most ROIs are very small. Also, there are deformable objects including animals (e.g. THIS SHAGGY DOG, THIS GUINEA PIG and THIS CHIHUAHUA) and THIS TURQUOISE STROLLER, with which local feature-based instance search does not deal easily. The final performance is evaluated by the MAP, where the mean is taken over all topics.

B. System Overview

The flowchart shown in Fig. 1 corresponds to our system created for TRECVID INS 2015. Given a topic and the collection of test videos, two full searches are individually conducted with BOVW and EWGR, resulting in two ranking lists RL1 and RL2, respectively. RL1 and RL2 are fused into RL3; a reranking step is performed with AF on the basis of RL1, resulting in RL4. Finally, RL3 and RL4 are fused into RL5, which corresponds to our submissions F_A_NTT_1 and F_A_NTT_2. F_A_NTT_1 and F_A_NTT_2 are basically the same with the exception of their configurations, which are explained in detail in Section V. F_A_NTT_3 corresponds to a fusion between RL1 and RL4. F_A_NTT_4 is exactly the run returning RL3. For the final

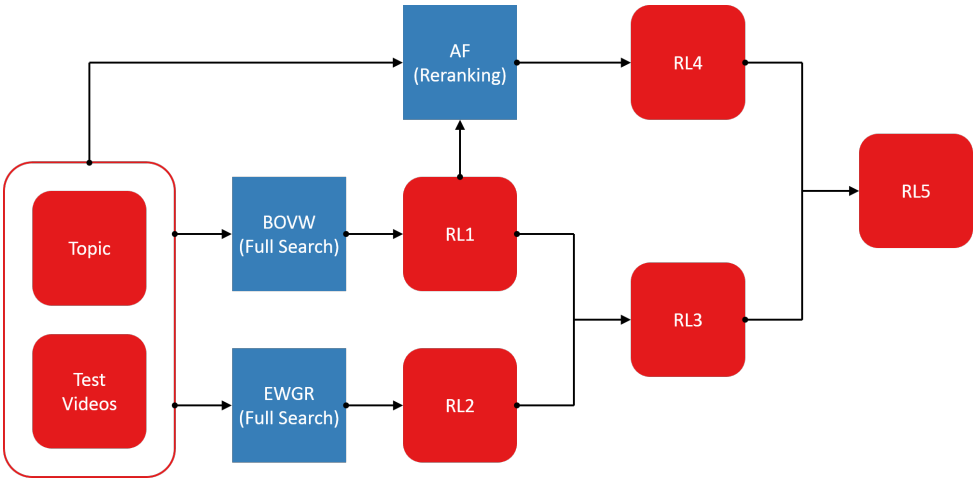

Figure 1: *Flowchart of our system created for TRECVID INS 2015. RL stands for* RANKING LIST. *RL5 corresponds to our submissions F_A_NTT_1 and F_A_NTT_2.*

submissions to TRECVID INS 2015, we first tested our system on the topics of INS 2014 for parameter tuning, and then employed the tuned system on the topics of INS 2015.

III. BAG OF VISUAL WORDS

A. Method

The BOVW model [11, 15] of local features has been shown to be successful in the instance search of near-rigid objects. When an image is represented using BOVW, it can be treated as a document. BOVW includes several steps for defining a set of visual words in this document: feature detection, feature description and vector quantization. After feature detection, each image is abstracted by a set of local patches known as interest points. Feature description methods such as SIFT [6] represent the interest points as numerical vectors known as local features or local feature descriptors. The final step is clustering, e.g. with approximated k-means [9, 11], over all local features to produce a visual vocabulary. Prior to the end of the clustering, an assignment step such as k-nearest neighbor is performed to associate each interest point with a visual word. The image can thus be represented by a set or a histogram of the visual words.

B. Implementation

Our implementation of BOVW follows the standard framework proposed and used in the literature [11, 12, 18]. More strictly, we first detected interest points with a Hessian affine region detector [7] from all the keyframes and extracted local feature descriptors (SIFT) [6]. In total, more than 15 billion SIFTs were extracted and converted to Root SIFTs [1]. A large visual vocabulary made up of one million visual words was then trained by an approximate k-means [9, 11]. Instead of taking k-means over the whole collection of test videos, we randomly sampled 100 million of the 15 billion features and include them in the clustering step. During the assignment of visual words, each feature can be assigned to a number k of nearest visual words in the context of soft assignment [12]. In our system, we selected $k = 3$ for the images of all topics and $k = 1$ for the keyframes of all test videos. In other words, we chose to perform hard assignment for test videos instead of soft assignment to avoid the large computational burden of a full search. These preprocessing steps are summarized in Fig. 2.

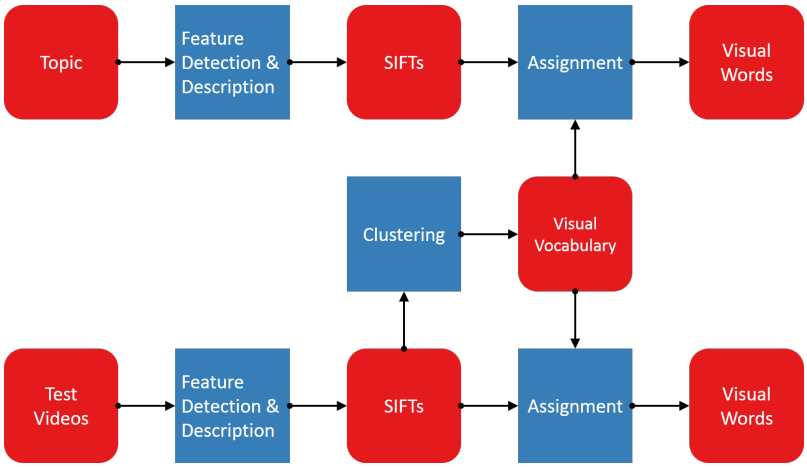

Figure 2: *Preprocessing of BOVW and EWGR.*

Table 1: *Performance of instance search based on BOVW.* TIME *excludes I/O time and the time taken for feature extraction and ranking, and is in units of second per topic.*

Configuration	σ^2 [12]	**MAP (2014)**	**MAP (2015)**	**Time (2014)**
1	1	26.0	–	3.73
2	.1	26.3	–	3.87
3	.01	27.4	28.4	3.88
4	.001	27.3	–	3.45

After following the above steps, we have a very large collection of 9.8 million keyframes, each of which is represented by a set consisting of 1546.6 local features on average. Each local feature is associated with a visual word as well as a set of affine parameters deriving from the Hessian affine region detector. Note that this collection is also used as the basis for EWGR as described in Section IV-A.

In BOVW, each keyframe was further encoded into a 1M-dimensional term frequency-inverse document frequency (TFIDF) histogram. The cosine similarity between each image in the topic collection and each keyframe in the test collection was efficiently computed with an inverted index. In practice, both the ROI and the background region inside the topic images contribute to the instance search. Therefore, we encoded each topic image into an ROI and a non-ROI TFIDF histogram and computed two similarities. The similarities were combined with a weighted average in which the ROI and non-ROI weights were 0.9 and 0.1 respectively. Suppose that a topic contains n images and a master shot contains m keyframes. An average pooling scheme was used to combine all of the $n \times m$ similarities, resulting in the topic and the master shot having one single similarity. The ranking list RL1 in Fig. 1 can thus be created by sorting all the master shots in descending order of this similarity.

C. Experimental Result

The performance of our instance search based on BOVW is summarized in Table 1. Our soft assignment scheme is in accordance with Philbin et al.'s study [12], in which the term frequency (TF) was computed by accumulating the weights assigned to each feature in a topic image. This weight is an exponential function of the distance from the feature descriptor to the cluster center of the k-nearest ($k = 3$) visual words: $\exp(-d^2/(2\sigma^2))$, where d is the distance and σ is a free parameter.

IV. Ensemble of Weak Geometric Relations

A. Preliminaries

After following the steps described in Section III-B, we have a very large collection of 9.8 million keyframes, each of which is represented by a set consisting of 1546.6 local features on average. Each local feature is associated with a visual word as well as a set of affine parameters derived from the Hessian affine region detector. An image is represented by a set P of features. For each feature $p \in P$ we are given its visual word $u(p)$, position $\mathbf{t}(p) = [x(p)\ y(p)]^{\mathrm{T}}$, scale $\sigma(p)$ and orientation $\mathbf{R}(p)$. p can be mapped, from a unit circle heading a reference orientation, by a 3×3 transformation matrix $\mathbf{F}(p)$:

$$\mathbf{F}(p) = \begin{bmatrix} \mathbf{M}(p) & \mathbf{t}(p) \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{1}$$

where $\mathbf{M}(p) = \sigma(p)\mathbf{R}(p)$ is a linear transformation and homogeneous coordinates are to be used for the mapping. In our system, $\sigma(p)$ is given by a real scalar and $\mathbf{F}(p)$ specifies a similarity transformation. $\mathbf{R}(p)$ is an orthogonal 2×2 matrix with $\det \mathbf{R}(p) = 1$, represented by an angle $\theta(p)$. Given two images P and Q, a correspondence $c \triangleq (p, q)$ is a pair of features $p \in P$ and $q \in Q$ with $u(p) = u(q)$. In EWGR, we assume $|C| \geq 2$ with $C = \{c\}$ and:

$$c = \big(u(c), \mathbf{t}(p), \sigma(p), \theta(p), \mathbf{t}(q), \sigma(q), \theta(q)\big). \tag{2}$$

B. Problem Formulation

We focus on the Cartesian product $C^2 = C \times C$, i.e. the set of all ordered pairs (c_a, c_b) where $c_a, c_b \in C$. A constraint function $h : C^2 \rightarrow \{0, 1\}$ is defined, which maps any arbitrary (c_a, c_b) to one if a given geometric constraint is satisfied, and zero otherwise. The geometric relation G is thus a subset of C^2 such that $\forall (c_a, c_b) \in G$, $h(c_a, c_b) = 1$. Accordingly, the spatial similarity can be formulated by the cardinality of G. Instead of a single constraint h, we build a set $H = \{h\}$ of geometric constraints, resulting in a new definition of G as in Eq. 3. Each $h \in H$ should be flexible as regards feature detection errors, but is allowed to offer a limited discriminative power. A conjunctive ensemble of such constraints creates a single strong constraint that is expected to be highly discriminating in terms of mismatches. The spatial similarity thus becomes $|G|$.

$$G = \left\{ (c_a, c_b) \in C^2 \,\middle|\, \left(\prod_{h \in H} h(c_a, c_b) \right) = 1 \right\} \tag{3}$$

The instance search based on EWGR is summarized in Fig. 3. The spatial similarity $S_{\mathrm{EWGR}} = |G|$ is then combined with the cosine similarity S_{BOVW} for the TFIDF histograms computed in Section III-B:

$$S_{\mathrm{RL3}}(P, Q) = \begin{cases} 1 + S_{\mathrm{EWGR}}(P, Q) & \text{if } S_{\mathrm{EWGR}}(P, Q) \neq 0 \\ S_{\mathrm{BOVW}}(P, Q) & \text{else.} \end{cases} \tag{4}$$

where S_{RL3} is the overall similarity. Equation 4 is the equivalent of first ranking the results according to S_{EWGR} and then ranking those with zero similarities via S_{BOVW}. The ranking list RL3 in Fig. 1 can thus be created by sorting all the master shots in descending order of S_{RL3}.

C. Geometric Relations

We focus on several fundamental classes of geometric coherence. Specifically, these classes include spatial neighborhood coherence, scaling coherence, rotation coherence, and relative position coherence in a polar and

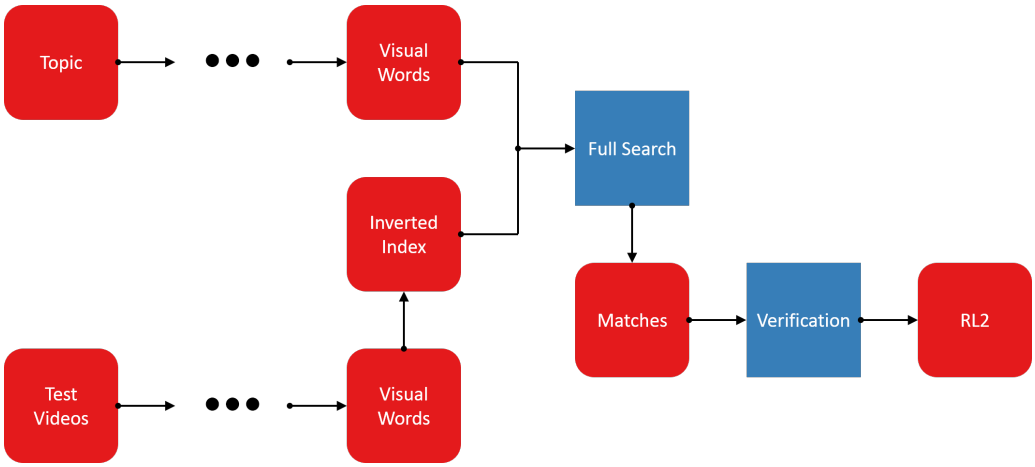

Figure 3: *Instance search based on EWGR.*

Table 2: *Performance of instance search based on EWGR.*

Configuration	k [16]	ϵ_θ [16]	$\epsilon_{\mathbf{v}}$ [16]	**MAP (2014)**	**MAP (2015)**
1	50	$\pi/8$	5	28.74	–
2	50	$\pi/8$	4	28.66	–
3	50	$\pi/8$	3	28.59	–
4	50	$\pi/8$	2	29.18	–
5	50	$\pi/8$	1	29.51	–
6	80	$\pi/8$	5	28.74	–
7	80	$\pi/8$	4	28.75	–
8	80	$\pi/8$	3	28.84	–
9	80	$\pi/8$	2	29.25	–
10	80	$\pi/8$	1	29.58	29.94

a Cartesian coordinate system, respectively. The conjunctive ensemble of these geometric constraints provides a high discriminative power in terms of mismatches. More detail on the definition of these constraints and their relationship with classic spatial verification methods can be found in our previous study [16].

D. Experimental Result

The performance of our instance search based on EWGR is summarized in Table 2. It depends on three parameters: k used for the spatial neighborhood coherence, $\epsilon_\theta \in (0, \pi)$ used for the rotation coherence and the relative position coherence in a polar coordinate system and $\epsilon_{\mathbf{v}} \in \mathbb{R}^+$ used for the relative position coherence in a Cartesian coordinate system. Among various configurations, we chose to show only the relationship between MAP and $\epsilon_{\mathbf{v}}$, because EWGR was the most sensitive to this parameter in our experiments. A smaller $\epsilon_{\mathbf{v}}$ corresponds to a stronger constraint on relative position coherence. As shown in Table 2, EWGR performed better when a stronger constraint on relative position coherence was imposed on the correspondences.

By comparing the MAPs in Table 2 with their counterparts in Table 1, we can see that BOVW without spatial verification was consistently outperformed by EWGR. The processing time on a per topic basis was 31.5 minutes and 27.0 minutes for INS 2014 and INS 2015, respectively. Because EWGR can be easily

processed in parallel, the processing time on a per topic image basis will be around 24 seconds and 20 seconds with 20 CPUs, which moves closer towards the realization of real-life applications. It should be noted that EWGR searched the full database containing around 9.8 million images.

V. Angle Free

A. Method

Suppose that two images are related as regards a common near-rigid object and an unknown affine transformation $\mathbf{A}$. $\mathbf{A}$ can be decomposed as a camera motion [8] as

$$\mathbf{A} = \Delta s \begin{bmatrix} \cos(\Delta\psi) & -\sin(\Delta\psi) \\ \sin(\Delta\psi) & \cos(\Delta\psi) \end{bmatrix} \begin{bmatrix} 1/\cos(\Delta\theta) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\Delta\varphi) & -\sin(\Delta\varphi) \\ \sin(\Delta\varphi) & \cos(\Delta\varphi) \end{bmatrix}, \tag{5}$$

where $s > 0$ is the camera zoom, ψ is the camera spin, and θ and φ parameterize the viewpoint angles. Δs indicates the scaling factor between two zooms, and $\Delta\cdot$ for the other three parameters indicates the difference between two camera angles. This decomposition suggests that we can simulate view-directional changes by varying the two camera angle parameters θ and φ, and can generate affine transformed images. After detecting local features from each of such simulated images, we append the simulated angle parameters θ and φ to all of the local features. We do not consider simulations with s and ψ because the two parameters can be obtained from the scale and the orientation of each local feature. After appending s and ψ, each local feature can be represented as $\mathbf{L} = \{s, \psi, \theta, \varphi, x, y\}$. Note that θ and φ depend on the simulations and are identical for all local features in the same image, while s and ψ are different for each local feature.

This simulation enables us to realize sufficient matching of local features under large viewpoint changes, but magnifies the negative impact of mismatches. The accuracy can be further improved by imposing view-directional consistency constraints on matches. In AF, we choose Hough voting as the solution because its time complexity is linear as regards the number of matches.

For each match, we can estimate a 3D rotation $(\Delta\psi, \Delta\theta, \Delta\varphi)$, a scaling factor Δs and a 2D translation $(\Delta x, \Delta y)$, which specify a relative camera pose difference or a camera motion. Voting all matches onto a Hough space according to the camera motions identifies clusters of matches. Each cluster can be treated as a hypothesis of $\mathbf{A}$, in which the matches serve as hypothetical inliers. A cluster with a larger number of matches provides more evidence for its hypothesis, while a cluster with only a single match is more likely to indicate an incorrect hypothesis. In practice, we do not consider the 2D translation because the relative displacement of local features is not consistent on non-planar objects. We thus have a 4D voting map. The grid interval is empirically set at 2 for Δs, $\pi/6$ for $\Delta\psi \in [0, 2\pi]$, $\pi/12$ for $\Delta\theta \in [-\pi, \pi]$ and $\pi/15$ for $\Delta\varphi \in [-\pi/2, \pi/2]$. To avoid the problem of boundary effects, each match votes for the two closest grids.

B. Implementation

Instead of a full search, our system based on AF aims at reranking only a given number of top results sorted in the image-level ranking lists of RL1. The method first extracts SIFTs [6] with the difference of Gaussians (DOG) and their descriptors from the simulated topic and test images. For test images, we chose to focus on the simulation with θ only (i.e. to ignore the rotation with φ) for higher efficiency. Subsequently, local feature matching is performed between each topic image and each image in the test collection. Matches with small similarities are rejected by means of a ratio test [6]. Hough voting based on view-directional consistency constraints is conducted, where each match votes for a corresponding grid on the 4D voting map according to the difference of $\mathbf{L}$. All votes are updated in the form of:

$$V(\Delta s, \Delta\psi, \Delta\theta, \Delta\varphi) \leftarrow V(\Delta s, \Delta\psi, \Delta\theta, \Delta\varphi) + 1, \tag{6}$$

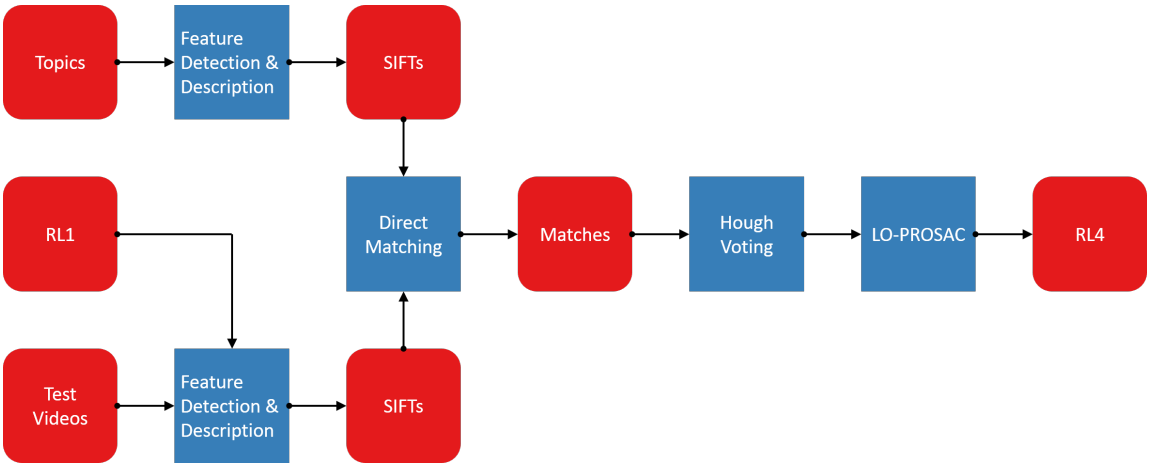

Figure 4: *Instance search reranking based on AF.*

Table 3: *Performance of instance search based on AF (INS 2014).* FUSION *indicates whether AF is late-fused with EWGR.* ROI WEIGHT *indicates the weights used for averaging the similarities obtained with the ROI only and with the full topic image.*

Configuration	Fusion	ROI Weight		Number of Top Results for Reranking				
		ROI	Full	1,000	2,000	3,000	6,000	10,000
1	–	1	0	28.65	28.93	29.08	29.46	–
2	EWGR	1	0	–	–	30.50	30.76	30.84
3	–	1	1	–	–	30.56	30.77	–
4	EWGR	1	1	–	–	31.64	31.78	31.49
5	–	0	1	28.73	29.86	30.10	30.14	–
6	EWGR	0	1	–	–	31.46	31.20	30.67

where $V(\cdot,\cdot,\cdot,\cdot)$ is the number of votes in the corresponding grid. The clusters identified by Hough voting are rejected if they contain fewer than seven entries. Each surviving cluster further undergoes a full geometric verification to achieve more accurate verification. We adopt LO-PROSAC [3,4] for this purpose, which takes epipolar geometry into account to discard outliers. Again, the clusters are rejected if they contain fewer than seven entries after LO-PROSAC. The final decision on the acceptance or rejection of a hypothesis is performed on the basis of Lowe's probabilistic model [5]. The similarity is defined as the total number of surviving inliers:

$$S_{\mathrm{AF}} = \sum_{}^{N} g(\Delta s, \Delta\psi, \Delta\theta, \Delta\varphi), \tag{7}$$

where N is the total number of grids within the Hough space and $g(\cdot,\cdot,\cdot,\cdot)$ the number of surviving inliers. The ranking list RL4 in Fig. 1 can thus be created by sorting all the master shots in descending order of S_{AF}.

C. Experimental Result

The performance of our instance search based on AF is summarized in Table 3. We mainly tested three configurations (or parameters): 1. the late fusion of AF with EWGR; 2. the contributions of ROI and non-ROI;

Table 4: *Our INS 2015 submissions.* #RERANKING *indicates the number of top results used for reranking. All ROI weights are* $1:1$.

Run ID	BOVW	EWGR	AF	#Reranking		MAP	
				ROI	Full	INS 2014	INS 2015
F_A_NTT_1	√	√	√	10,000	3,000	32.12	31.73
F_A_NTT_2	√	√	√	6,000	6,000	31.78	33.10
F_A_NTT_3	√	–	√	10,000	3,000	–	31.56
F_A_NTT_4	√	√	–	–	–	29.58	29.94

3. the number of top results used for reranking. In general, the combination with EWGR improved the MAPs in all cases. The MAPs were the highest when the weights for the ROI and the full topic image were set at $1:1$, i.e. both the ROI and non-ROI regions of the topic image contribute to the instance search. By comparing Configurations 5 and 6 with Configurations 3 and 4, we can see that the MAPs can be improved by putting more weight on the ROI. When only the ROI was used for reranking, a larger number of top results led to a higher MAP. However, if the full image is used, we can observe a slight degradation of MAP as we enlarge the set of top results from 6,000 to 10,000 images. We believe this is because of the disadvantageous effect that background regions have on the instance search.

We further increased the number of reranked images for ROI and reduced the number of reranked images for non-ROI (Table 4). By comparing the MAPs of F_A_NTT_1 and F_A_NTT_2 obtained with INS 2014 and INS 2015, we can see that the effects of the number of reranked images are inconsistent and so seem topic-dependent. We shall return to this subject in the future. With INS 2015, the processing time of AF on a per topic (INS 2015) basis was 78.2 minutes and 52.3 minutes (excluding the processing time of EWGR) for F_A_NTT_1 and F_A_NTT_2, respectively. As a reference, the commensurate time of EWGR was 27.0 minutes, as described in Section IV-D. It is important to remember that EWGR searched the full database containing around 9.8 million images while our system based on AF aimed at reranking no more than 10,000 images. AF required much longer processing time because the number of images was enlarged by the transformation simulation and we had to perform matching between much more pairs of images. Another reason relates to the direct matching of SIFTs adopted in AF where we abstained from using the visual vocabulary to avoid large quantization errors. Apart from the efficiency issue, AF effectively helped our system recognize and localize near-rigid instances in scenes with large 3D viewpoint changes from the topic.

VI. CONCLUSION

In this report, we proposed two practical spatial verification methods called EWGR and AF for instance search from videos. These methods achieved significantly improved accuracy in two ways: 1. by imposing a conjunctive ensemble of multiple pairwise geometric constraints on SIFT correspondences; 2. by detecting ASIFT correspondences and employing Hough voting in a 3D camera motion space. Both EWGR and AF contribute to the rejection of mismatches in a local feature-based instance search. The former was better adapted for discriminating and fast spatial verification. It supports the full search of a very-large-scale video database with 9.8 million keyframes extracted from 464 hours of video. In contrast, AF helped our system recognize and localize near-rigid instances in scenes with large 3D viewpoint changes from the topic. It can successfully handle relevant results that slip down the ranking list due to viewpoint variation when BOVW or EWGR is used. The two methods were complementary and so their combination resulted in further improvement as regards the instance search task. We took the framework of BOVW [17, 18] as the baseline, and achieved a more than 16% increase in MAP on INS 2015 through the combination of EWGR and AF.

REFERENCES

[1] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[2] Yannis S. Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.

[3] Ondrej Chum and Jiri Matas. Matching with PROSAC - Progressive sample consensus. In *CVPR*, pages 220–226, 2005.

[4] Ondrej Chum, Jiri Matas, and Josef Kittler. Locally optimized RANSAC. In *DAGM*, pages 236–243, 2003.

[5] David G. Lowe. Local feature view clustering for 3D object recognition. In *CVPR*, pages 682–688, 2001.

[6] David G. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[8] Jean-Michel Morel and Guoshen Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sciences*, 2(2):438–469, 2009.

[9] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, pages 331–340, 2009.

[10] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, and Roeland Ordelman. TRECVID 2015 – An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[11] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[12] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[13] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Spatially-constrained similarity measure for large-scale object retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1229–1241, 2014.

[14] Jun Shimamura, Taiga Yoshida, Yukinobu Taniguchi, Hiroko Yabushita, Kyoko Sudo, and Kazuhiko Murasaki. The method based on view-directional consistency constraints for robust 3D object recognition. In *MVA*, pages 455–458, 2015.

[15] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[16] Xiaomeng Wu and Kunio Kashino. Robust spatial matching as ensemble of weak geometric relations. In *BMVC*, pages 25.1–25.12, 2015.

[17] Xiao Zhou, Cai-Zhi Zhu, Qiang Zhu, Shin'ichi Satoh, and Yu-tang Guo. A practical spatial re-ranking method for instance search from videos. In *ICIP*, pages 3008–3012, 2014.

[18] Cai-Zhi Zhu, Yinqiang Zheng, Ichiro Ide, Shin'ichi Satoh, and Kazuya Takeda. Nagoya University at TRECVID 2014: The instance search task. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.