Nagoya University at TRECVID 2014: the Instance Search Task

Cai-Zhi Zhu[1] Yinqiang Zheng[2] Ichiro Ide[1]
Shin'ichi Satoh[2] Kazuya Takeda[1]

[1] Nagoya University, 1 Furo-Cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
[2] National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract. This paper presents our recent progress on a video object retrieval system that participated in the Instance Search (INS) task of the TRECVID 2014. Basically the system is a further extension of our previous Bag-of-Words (BOW) framework, with emphasis on pursuing a practical spatial re-ranking method scalable to large video database this year. We take the asymmetrical dissimilarities based system, which performed best in the INS2013 task, as the baseline, and re-rank with an improved spatial verification method. Experiments carried out the TRECVID INS2013 and INS2014 consistently show that, our re-ranking algorithm is able to further improve the baseline system at a rather fast speed.

Keywords: Spatial re-ranking, RANSAC, object retrieval, instance search, video retrieval.

1 Introduction

We address the problem of instance search (or visual object retrieval, equivalently) from videos, *i.e.*, to rank database videos according to the probability of the existence of specific objects delimited by regions in query images, as defined in the TRECVID INS task [7].

Our former work for this task mainly focused on how to fuse contributions from the foreground and background region of queries [17], how to design asymmetrical dissimilarities more suitable for measuring the existence of a small query ROI in a big database image [16], and how to aggregate multiple images/frames [17, 15]. In this paper, we study the spatial re-ranking for the video instance search purpose. As is well known, spatial re-ranking has proved to be successful in image retrieval [11, 10, 13]. Yet no work on spatial re-ranking has been systematically reported for instance search from videos so far: As videos are composed of multiple frames, and frame-by-frame spatial verification is too prohibitive, till now we lack an efficient spatial re-ranking method designed for large-scale video instance search. The effectiveness is unclear as well. Recently Zhang and Ngo [12] proposed a new spatial re-ranking method integrated with topology checking and context modeling. While one of main conclusions in their

work is that traditional RANSAC based spatial re-ranking methods do not necessarily work in the context of video instance search, which is also consistent with our preliminary nding in this aspect [16].

Here we challenge the previous conclusion that the general RANSAC based spatial re-ranking algorithm does not work for the video instance search purpose [14]. Based on our former Bag-of-Words (BoW) [11] framework [17, 16], we aim at improving the traditional spatial re-ranking method for the video instance search purpose, and nally come up with a practical re-ranking method with large improvement on both speed and accuracy. Experiments carried out on the INS2014 dataset shows that, the proposed re-ranking method signi cantly improves our baseline system, at an acceptable time cost.

Contributions: We rst make contributions to the speed: First, based on the available Bag-of-Words representations of frames, representative image/frame (of a query topic /database video) can be e ciently selected for later spatial veri cation, thus avoid the expensive process of verifying every frame extracted from videos. Second, in the process of building tentative matching feature pairs as the input to the RANSAC algorithm, features quantized to the same visual word are regarded as tentative matches by leveraging the VQ information, thus fully avoid the expensive Nearest-Neighbor (NN) search. These two steps lead a signi cant speedup. For instance, it costs less than 0.05 second to verify a pair of query topic and video clip on average. Another contributions lie on the performance. we come up with a ROI-originated RANSAC method, of which the geometrical transformation matrix is estimated from the ROI, while the number of inliers used for re-ranking is computed on the full image.

The rest of this paper is organized as follows. Section 2 introduces the TRECVID benchmark dataset, and the BoW baseline system that serves as the front-end of the proposed re-ranking method. Section 3 proposes our practical spatial re-ranking method comprised of several improvements; experiments and analysis for each improvement are also given. We brie y introduce our INS2014 submissions in Section 4. Finally we conclude the paper in Section 5.

2 Benchmark and Baseline

2.1 The TRECVID INS2013INS2014 Dataset

The TRECVID INS Datasets are rather challenging because they were designed to simulate real demands for the multimedia retrieval community. The task description of TRECVID instance search challenge is as follows [8]: given a collection of test video clips and a collection of query topics that delimit a person, object, or place, for each query topic, up to one thousand video clips most likely to contain a recognizable instance of the query topic should be returned.

INS2014 and INS2013 use the same database, $i.e.$ BBC EastEnders videos. Their only di erence is the query topics, as shown in Figure 1. There are 469,539 videos with total duration around 430 hours in the database. Each video clip last 3$\sim$4 seconds on average. We sampled frames at 5fps and extracted 7,802,853

(a)

(b)

Fig. 1: A random sampling of images (Programme material copyrighted by BBC) from all 30 query topics in: (a) INS2013 and (b) INS2014. Red shapes that delimit ROIs are overlaid.

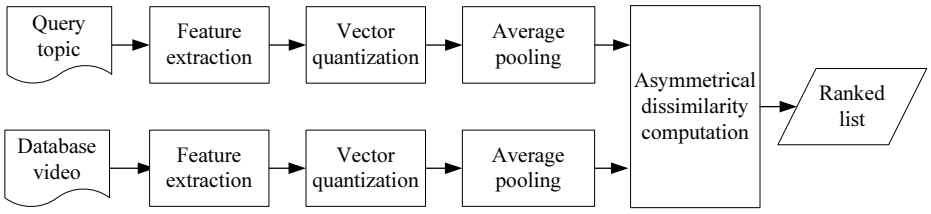

Fig. 2: The baseline system.

frames in total. For each of the INS dataset, there are 30 query topics with each consists of four images. A random sampling of images of all 30+30 query topics are shown in Figure 1, in which red shapes that delimit ROIs are overlaid. We can observe that the query topics are very challenging, as most query objects are very small and non-rigid objects exist. The nal performance is evaluated by the o cial single score, i.e., mean inferred average precision (infAP) , where the mean is taken over all query topics. Note in this paper the average precision is reported in percent accuracy.

2.2 The Baseline System

We take the following BoW system shown in Figure 2 as the baseline system, with two important optimizations: First, the average pooling method, proposed by Zhu and Satoh [17], was adopted to aggregate BoW vectors of multiple images of each query topic and multiple frames of each database video [15]. Second, asymmetrical dissimilarities (by Zhu *et al.* [16]), instead of symmetrical ℓ_p distances, were computed as the ranking score.

For the TRECVID INS2014 submission, we rst test our system on the INS2013 dataset for parameter tuning, and then apply the well-tuned algorithm on the new INS2014 dataset for nal submission. The nal submission is the fusion of results returned by six di erent SIFT features, including Hessian-A ne, Harris-Laplace, MSER three detectors, and color SIFT and Root-SIFT [1] two descriptors. We found that the fusion of multiple features only brings marginal improvement over the best single feature { Hessian-A ne Root-SIFT, while at the sixfold time cost. Therefore, to make it simple, next we will report our experiment on the single SIFT feature and on one dataset, while conclusions could be generalized to the fused system and the other dataset.

To be more speci c, we rst detected a ne-Hessian interest points [6] from every video frame and query image and extracted SIFT descriptors. In total around 12 billion SIFT features were extracted and converted to RootSIFT [1] hereafter. Then a large visual vocabulary made up of one million visual words was trained by an e cient approximate k-means algorithm (AKM) [9]. Hereafter each image was encoded into a one-million dimensional term frequency-inverse document frequency (tf-idf) vector, by quantizing each SIFT descriptor with the vocabulary. An average pooling scheme [17] was used to aggregate all BoW

Table 1: Comparison with the baseline and different configurations of the spatial re-ranking method on the INS2013 dataset.

Method	Query	Video	Match	CReg	SReg	infAP	Mins
M1	RMD	RMD	VQ1	ROI	FUL	31.61	19
M2	REP	REP	VQ1	ROI	FUL	33.49	20
M3	ALL	ALL	VQ1	ROI	FUL	34.29	1,440
M4	ALL	REP	VQ1	ROI	FUL	33.77	80
M5	ALL	REP	VQ2	ROI	FUL	34.39	96
M6	ALL	REP	VQ3	ROI	FUL	34.58	128
M7	ALL	REP	NN	ROI	ROI	30.28	800
M8	ALL	REP	VQ1	ROI	ROI	33.72	80
M9	ALL	REP	VQ1	FUL	ROI	33.32	200
M10	ALL	REP	VQ1	FUL	FUL	28.00	200
BL	–	–	–	–	–	31.33	–

Note: RMD, REP and ALL in the column Query/Video denote spatial verification with a random selection of image/frame, a representative image/frame (Subsection 3.1), and all images/frames, respectively. VQk ($k = 1, 2, 3$) and NN in the column Match denote building tentative matches by the proposed VQ method (Subsection 3.2) and the NN method, respectively. CReg and SReg denote the region in which SIFT features are used to compute the transformation matrix and to determine the number of inliers for scoring, respectively. ROI and FUL in these two region columns denote query ROI and full query image(Subsection 3.3), respectively. Time is reported as the total mins of re-ranking all 30 query topics on the top 1k list, *i.e.*, 30*1,000 query topic and database video pairs. BL in the Method column is the baseline method (*i.e.*, the δ_1 method [16]) described in Subsection 2.2, of which the output is used for re-ranking.

vectors, of images contained in a video or a query topic, into one vector to represent that video or query topic. Finally, the asymmetrical dissimilarity score [16], between every pair of query topic and database video, was efficiently computed by an inverted file index.

The result of this optimized baseline system, which won the TRECVID INS2013 task [8], is taken as the initial ranked list for further spatial re-ranking.

3 A Practical Spatial Re-ranking Method, Experiments and Analysis

As the spatial re-ranking itself is computationally expensive, so far a common strategy reached is to re-rank only a limited number, *e.g.*, one thousand as in our experiments, of images on the top of the initial ranked list. Note the speed issue becomes even severe in the case of video search. As videos are composed of multiple frames, and to spatially verify every frame is too prohibitive, till now we lack an efficient spatial re-ranking method designed for the video instance search purpose. The effectiveness is unclear as well.

Fig. 3: Comparison of ranked lists on INS2013. Query objects are on the left side. On the right side, the top 10 returns are ranked from left to right: For each example, the upper and lower rows are returned by BL and M6 in Table 1, and the accuracies from top to bottom are 9.46 *vs.* 25.84, 20.71 *vs.* 32.37, and 31.21 *vs.* 65.35. Positive (negative) samples are marked with green (red) bounding boxes. Programme material copyrighted by BBC.

In this section, we introduce three improvements on the traditional spatial re-ranking method in three sub-sections, respectively. We also analyze the rationality of each part by comparison experiments, as shown in Table 1. Figure 3 illustrate the advantage of our proposed re-ranking method by comparing it with the baseline in Subsection 2.2.

3.1 Representative Image/Frame Selection

As we mentioned before, in the TRECVID INS dataset, each database video contains multiple frames, and each query topic comprises of several query images. This necessitates the selection of a limited number of representative pairs, of video frame and query image, for spatial veri cation, as to verify every frame-image pair is too prohibitive. We propose an e cient method that regards the frame-image pair whose BoW histogram intersection is maximal as representative, denoted by (i_{rep}, j_{rep}), as described in Eq. 1.

$$(i_{rep}, j_{rep}) = \arg \max_{i \in \mathcal{I}, j \in \mathcal{J}} \|\min(\mathsf{Q}_i, \mathsf{V}_j)\|_{\ell_1}, \tag{1}$$

here Q_i, V_j are the BoW vector of query image i and video frame j, respectively. $\mathcal{I}, \mathcal{J}$ are the set of images contained in that query topic and set of frames in that video, respectively. i_{rep} and j_{rep} are the representative query image and video frame, respectively. If we can a ord verifying all query images, we can also choose to select representative frame j_{rep_i} for every query image $i (i \in \mathcal{I})$, as in Eq. 2.

$$j_{rep_i} = \arg \max_{j \in \mathcal{J}} \|\min(\mathsf{Q}_i, \mathsf{V}_j)\|_{\ell_1}, \forall i \in \mathcal{I}, \tag{2}$$

Experiments and Analysis: By comparing M2 (Eq. 1) and M1 in Table 1, we can see that the proposed representative selection method signi cantly outperforms the random method. The M2 method that selects only one image-frame

pair for spatial verification is only slightly inferior to the M4 method (Eq. 2) that verifies all query image and their corresponding representative frame, and also the M3 method that exhaustively verifies all image-frame pairs. Note M3 is actually not practical in most cases due to the speed. For M3 and M4, we summed up the number of all inliers, detected in each individual process of matching a single image-frame pair, as the final re-ranking score.

3.2 VQ Based Tentative Feature Matching

The RANSAC algorithm requires tentative matching feature pairs as input. Generally the feature matching process is as follows [5]. The first and second (approximate) NN of each query SIFT will be retrieved first, and then the ratio of their distances from the query SIFT is tested against a threshold, $e.g.$, 0.8 as in our experiments (M7 in Table 1). The first NN will be accepted as a tentative match only if it is close enough to the query SIFT. The above process has two problems. One is that raw SIFT features have to be kept in storage, which usually consume huge disk space (on the order of terabyte) for a large dataset, such as the TRECVID INS2013. The other is that (approximate) NN search is computationally expensive, especially for high dimensional feature vectors, and verifying a large number of images (which is usually necessary for better accuracy) becomes impractical.

To tackle this issue, we propose a VQ based feature matching method, in which features quantized to the same visual words are considered as matches. Assume each SIFT feature will be assigned to k number of nearest visual words in the context of soft assignment [10, 4], a pair of features will be regarded as a tentative matching pair as long as they have at least one common nearest visual word. As our front-end system is based on the BoW framework, the VQ information is directly available without extra computation, thus the proposed method is much faster than the NN based method. Another advantage is that the VQ based matching method also avoids the necessity of storing and accessing raw SIFT features on the disk.

Experiments and Analysis: Comparing M7 and M8, we can see that the NN based matching method is remarkably worse than the VQ based method. That is because the VQ method use a large vocabulary (one million in our experiments) as the reference to judge matching. In contrast, the NN based method relies on the relative distance of NNs in a set of features extracted from a video. The ratio of distance is not always inaccurate due to the burstiness effect [3], and also when the feature set is small [18].

For the VQ based method, the relationship between the performance and $k (k \in [1, 3])$ is reflected by M4-6 in Table 1. We can see that a large k increases the performance, that is because the number of consensus matching feature pairs will increase. The RANSAC algorithm we used is very good at discovering those consensus pairs, though the number of false matching pairs increases as well. We also observed that the NN based method is slower than the VQ based method by almost one order of magnitude.

3.3 ROI-Originated RANSAC

The query ROI information was once used by Zhu etc. [16] to design asymmetrical dissimilarities that outperform their symmetrical counterparts. Here we further leverage the ROI information to design a ROI-originated RANSAC method, in which a similarity transformation matrix [9] will be first calculated from SIFTs located in the query ROI, thus we obtain a object-focused, therefore more accurate, transformation matrix. Hereafter, SIFTs of the full query image will be verified by the matrix, and the number of all inliers, including those inside and outside the ROI, will be used for re-ranking.

Experiments and Analysis: The column CReg and SReg in Table 1 compare two different options of regions used for computing the transformation matrix and scoring. Comparing M4 and M8-10, we can see that the transformation matrix is better to be computed from the ROI. In this case the given ROI will act as a priori information and lead to a object-focused transformation matrix. In contrast, the verification should be performed on the full image, as the background is also very useful to identify positive samples, due to near duplicates exist in the TRECVID dataset.

4 Our INS2014 Submissions

Our INS2014 submissions on behalf of NU team can be summarized in Table 2, from which we can conclude: (1) the asymmetrical δ_2 method is consistently better than the δ_1; (2) Our proposed re-ranking method further improves the performance from 28.77 to 30.44, which is ranked second among all 22 teams.

As a new change in INS2014, for the same algorithm, participants are allowed to submit multiple runs with the number of query images ranging from 1 to 4 (corresponding to the query set ID: A to D, respectively), and videos from which query images were extracted could be used as the reference (ID: E). This way each team could submit maximally 5 runs for the same algorithm. Our submissions on behalf of NU team are summarized in Table 2. By transversely comparing the results returned by different query sets, we could draw the same conclusion as in our former research [15]: normally more query images better performance, while simply pooling all query images together might not be the best aggregation way. Note that in some cases the performance could even drop with more query images being used (compare F_D_NU_3 and F_E_NU_3 in the table).

5 Conclusions

In this paper, we propose a practical spatial re-ranking method for instance search from videos, which achieves significant speedup in two aspects, one is to select representative image/frame for re-ranking, the other is to use VQ information, instead of NN search with raw SIFTs, to build tentative matching feature

Table 2: Our submissions for the TRECVID INS2014.

Method ID	x=A	x=B	x=C	x=D	x=E	Method Description
F_x_NU_1	–	–	–	30.44	–	δ_2+re-ranking
F_x_NU_2	19.11	24.56	26.50	28.77	28.99	δ_2 [16]
F_x_NU_3	18.33	21.95	23.71	25.56	21.42	δ_1 [16]
F_x_NU_4	16.00	18.75	22.27	24.34	–	An implementation of HE [2]

Note: The official performance were reported on 27 query topics after three topics (9100, 9113 and 9117) being excluded.

pairs. These matching pairs will serve as the input to the process of geometrical transformation estimation. These two steps greatly reduce the computational burden via using the available BoW vectors and VQ information, and also avoid the necessity of storing and accessing raw SIFT features. We also propose a ROI-originated RANSAC method leveraging the known ROI information to compute the geometrical transformation matrix, while scoring by the number of all inliers detected on the full image. We took the best reported results as the baseline, and achieved around ten percent increase in performance on both INS2013 and INS2014 by the proposed method, at a much fast speed.

References

1. R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.
2. H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.
3. H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.
4. H. Jégou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007.
5. D. Lowe. Distinctive image features from scale-invariant key points. *IJCV*, 60:91–110, 2004.
6. K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1:63–86, 2004.
7. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.
8. P. Over, J. Fiscus, G. Sanders, B. Shaw, G. Awad, M. Michel, A. Smeaton, W. Kraaij, and G. Quénot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013*, 2013.
9. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.
10. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

11. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.
12. W. Zhang and C.-W. Ngo. Searching visual instances with topology checking and context modeling. In *ICMR*, 2013.
13. W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, 2010.
14. X. Zhou, C.-Z. Zhu, Q. Zhu, Y.-T. Guo, and S. Satoh. A practical spatial re-ranking method for instance search from videos. In *ICIP*, 2014.
15. C.-Z. Zhu, Y.-H. Huang, and S. Satoh. Multi-image aggregation for better visual object retrieval. In *ICASSP*, 2014.
16. C.-Z. Zhu, H. Jégou, and S. Satoh. Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *ICCV*, 2013.
17. C.-Z. Zhu and S. Satoh. Large vocabulary quantization for searching instances from videos. In *ICMR*, 2012.
18. C.-Z. Zhu, X. Zhou, and S. Satoh. Bag-of-words against nearest-neighbor search for visual object retrieval. In *ACPR*, 2013.