Discriminative Reranking based Video Object Retrieval

Santhoshkumar Sunderrajan*, Niloufar Pourian*, Mahmudul Hasan**, Yingying Zhu***,
B.S.Manjunath* and Amit Roy Chowdhury***
*Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA-93106
**Dept. of Computer Science and Engineering, University of California, Riverside, CA-92521
***Dept. of Electrical Engineering, University of California, Riverside, CA-92521

***Abstract*—For the instance search task, we are given a set of query images with the corresponding textual meta-data and objects masks to retrieve video shots containing query objects from FLICKR video database. We extract meaningful regions in the key-frames using Maximally Stable Extremal Regions (MSER) and use SIFT descriptors for representation. We use standard Bag of visual Word (BoW) model to represent database images. Additionally, we crawled training images for each query topic using the textual meta-data from Google and FLICKR images databases to train a discriminative classifier using Support Vector Machines (SVM). We use a discriminative model to re-rank candidate images obtained by initial BoW search. The experimental results demonstrates the efficacy of the overall system. Finally, we highlight the need for domain adaptation when the source and target domains are completely different.**

I. Introduction

The aim of the instance search task is to retrieve video shots containing a particular query topic that is specified by multiple images, their associated masks marking the area of interest, and a textual meta-data. Retrieving an object in a database of images is a challenging task because an object's visual appearance may completely vary due to the changes in viewpoint, scale changes, lighting, and occlusion. For this reason, region extraction and descriptors are required to be built with some degree of invariance to viewpoint and illumination conditions.

TRECVID-2012 Instance Search task (INS) is a pilot task that concentrates on evaluating several algorithms for video object instance retrieval [6]. Videos shots are created from FLICKR video database. Participants are given with twenty-one query topics to retrieve from the given video database. The given query images appear in one or more video shots and the task is to retrieve video shots that contain the object of interest.

In previous years, because of the missing labels in the testing dataset, most of the participants used Bag of visual Words model in combination with some form of nearest neighbour search. In contrast, we used a combination of unsupervised retrieval with a discriminative re-ranking strategy. For supervised training, we crawled training examples from Google and FLICKR image databases using the textual meta-data available with query topics.

The rest of the paper is organized as follows: Section II discusses the general framework for video object retrieval. Section III discusses about discriminative re-ranking for improving the retrieval task in an efficient manner. Section IV demonstrates the results of experiments and finally we conclude in section V

II. Video Object Retrieval with Bag of Visual Words

We follow the standard BoW retrieval framework described in [9]. We retrieve key frames for the video database and extract Maximally Stable Extremal Regions (MSER) [5] and describe the regions using SIFT [4]. Finally, we represent the images with Bag of Visual words model using the dictionary trained from the TRECVID 2011 Instance Search task dataset (BBC Videos). We retrieve similar images from the testing database using chi-square distance matching and finally we re-rank the candidate list using a discriminative classifier trained from an auxiliary dataset. Figure II shows the complete framework used for the instance search task. Rest of this section explains each of the above mentioned steps in detail and discriminative re-ranking is explained in section III.

A. Key-frame Extraction

We extracted key-frames in the training dataset (BBC videos from 2011 task) using the FFMPEG utility. For the test dataset, we sampled images every 15 frames and the test database consisted of $223,141$ key-frame images.

B. Region Extraction and Feature Descriptors

For every image in the training and testing databases, we extracted Maximally Stable Extremal Regions as described in [5]. These are the regions for which the area is approximately stationary as the intensity threshold is varied. We used SIFT descriptors to represent each of these extracted regions.

C. Codebook Generation

For generating codebook from the training images, we randomly chose one million SIFT descriptors extracted from various regions in the entire training database. We used approximate k-means clustering along with the stop list criteria to obtain the final codebook.

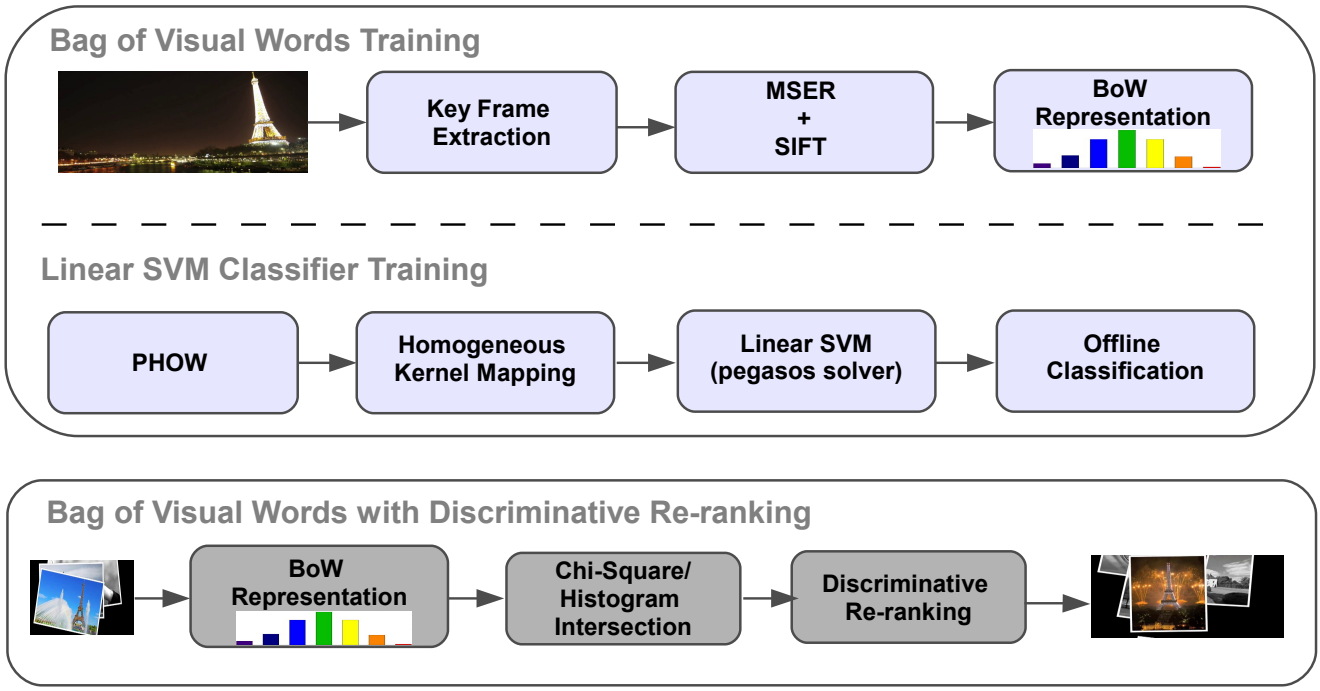

Fig. 1. Video object retrieval with discriminative re-ranking framework.

1) Approximate k-means Clustering: In typical k-means clustering algorithm, a great amount of computation time is spent in finding distances between the points and cluster centres. In approximate k-means clustering, an approximate neighbour method using k-d tree data structure is used [7]. The algorithm complexity of a single k-means iteration is reduced considerably. For our computation, we chose a larger dictionary size (K = $10,000$) for a better performance.

2) Stop List Criteria: Using a stop list strategy the most frequent words that occur in most of the images are discarded. These are the noisy features that do not provide any useful information. In addition to this, we also removed least frequently occurring features from the codebook list. For our experiments, we discarded top 5% and bottom 10% of the codebooks entries. We finally ended up with a codebook size of K = $8,500$.

D. Retrieval

We represented each image by a normalized histogram using the codebook obtained from the training dataset (BBC Videos). For a given query image we computed the BoW model using a similar strategy and compared with the database images using Chi-square and Histogram intersection metrics [3]. We ordered images based on the matching score. Since the query image and database images vary to a great extent, we performed a discriminative re-ranking to enhance the performance of the retrieval as discussed in section III.

III. Discriminative Re-ranking

In the given test dataset, the object might appear in arbitrary location and undergo arbitrary distortion and transformation when compared to the query images. Hence in order to fully capture the query object characteristics, we need some form of query expansion or re-ranking mechanism [1]. We adopt a discriminative re-ranking mechanism by modelling the query object characteristics explicitly using the images obtained from the internet in an offline manner.

A. Crawling Training Images

Using the textual meta-data available with the query images, we auto-crawled 200-300 images per topic from the Google and FLICKR image databases.

B. Discriminative Learning

With the images obtained from the web, we first extracted Dense SIFT for every image and then formed PHOW descriptors [2]. We encoded the PHOW descriptor using homogeneous kernel mapping. Finally, we trained a linear classifier "1 v/s all" classifier using SVM with the Pegasos solver [8]. We classified all the key frames using the model learned for query type and associated a likelihood score. We used this score to re-rank the candidate list obtained from BoW model based retrieval. Rest of this section explains each of the above mentioned steps in detail.

1) Feature Extraction: We extracted dense SIFT feature with a step size $= 4$ pixels i.e. the grid at which features are extracted. We used k-means clustering to generate a

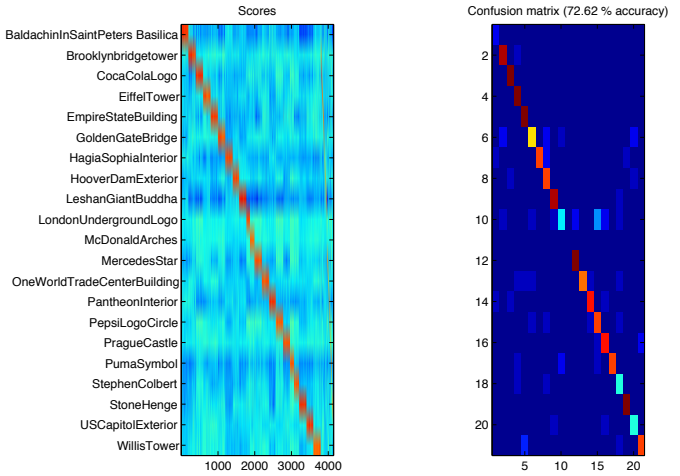

Fig. 2. Classification scores and the corresponding confusion matrix for different query topics obtained using SVM classifier on PHOW representation.

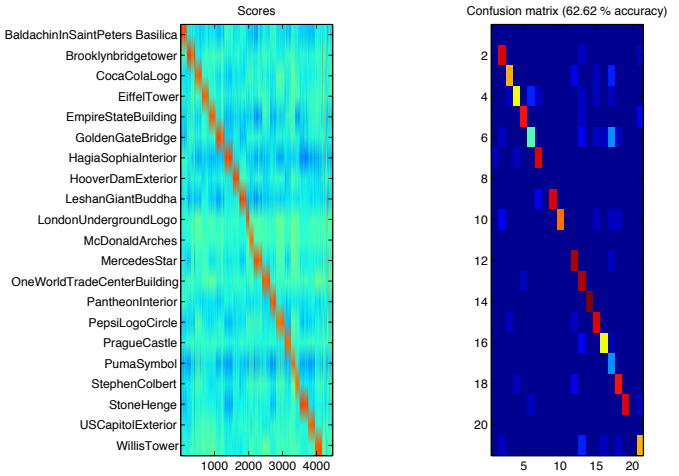

Fig. 3. Classification scores and the corresponding confusion matrix for different query topics obtained using SVM classifier on color based PHOW representation.

codebook of size C$=300$. The PHOW features are a variant of dense SIFT descriptors, extracted at multiple scales [2]. Additionally, for another set of experiments, a color version, named PHOW-color, we extracted descriptors on the three HSV image channels and stacked them together. For the color version, we used a step size of 7 pixels to extract dense SIFT feature.

2) Homogeneous Kernel Mapping: The homogeneous kernel map is a finite dimensional linear approximation of homogeneous kernels, including the intersection, chi-square, and Jensen-Shannon kernels [10]. These kernels are particularly useful for descriptors represented using histogram. For our experiments, we used a chi-square kernel.

3) Linear Support Vector Machine Classifier: We trained a linear SVM classifier using the Pegasos solver described in [8]. Pegasos is a simple and efficient iterative algorithm for the solving the optimization problem for SVM. The run-time of the solver does not depend on the size of the training dataset and hence it can be scaled to large datasets easily. We used chi-square kernel with homogeneity of kernel set to 0.5. In our experiments, we used two different classifiers trained on gray scale PHOW and a color version of it (HSV). Figures 2 and 3 show the classification scores and confusion matrix for different query topics for the two discriminative models trained from the internet images.

4) Off-line Classification: Since the model is trained by querying images from the internet using the textual meta-data, we reduced the runtime for retrieval by classifying each of the key frames using the model learned for all 21 query types in an off-line manner.

IV. Experiments

For TRECVID evaluation, we submitted three runs and each of the runs is discussed in detail in the following sub sections. Figure 4 shows number of hits per 1000 candidates retrieved for various runs compared to ground truth (gt) used for evaluation. As seen in figure 4, all three runs perform equally well for global location based queries such as $9051, 9052$ etc.

A. *Run-1: BoW with Chi-square distance*

For the first set of experiments, we used the Bag of visual words model with Chi-square distance for matching the given query images with key-frames in the test database and re-ranked using the classification scores obtained by SVM classifier trained on internet images. We combined the unique results from different sub-queries and ordered them based on the matching scores. Figure 5 shows the number of hits per 1000 candidates retrieved for run-1 versus the ground truth (gt) and the best result (best). Since the BoW is a global model, it is well suited for location based queries. We attribute the difference in performance compared to the best performance to the way in which the key-frame extraction is done. Since we extracted only 3 key-frames per video shot, we might have missed some object instances that appear for a small number of frames in a given video shot.

B. *Run 2: BoW with Chi-square distance + SVM on Color PHOW*

For the second set of experiments, we used the Bag of visual words model with Chi-square distance for matching the given query images with the key-frames in the database and re-ranked using the classification scores obtained by SVM classifier trained on internet images. We combined the unique results from different sub-queries for each topic and ordered them based on the matching scores. Figure 6 shows Average Precision for different queries (topics). Interestingly, one would expect the discriminative model trained on an exemplary dataset would perform well on the test dataset, however, due to inherent difference in feature distribution, learned model does not fair so well in the test dataset. Compared to run-1, for query 9048 (Mercedes star logo), run-2 performs well due to the context information used while

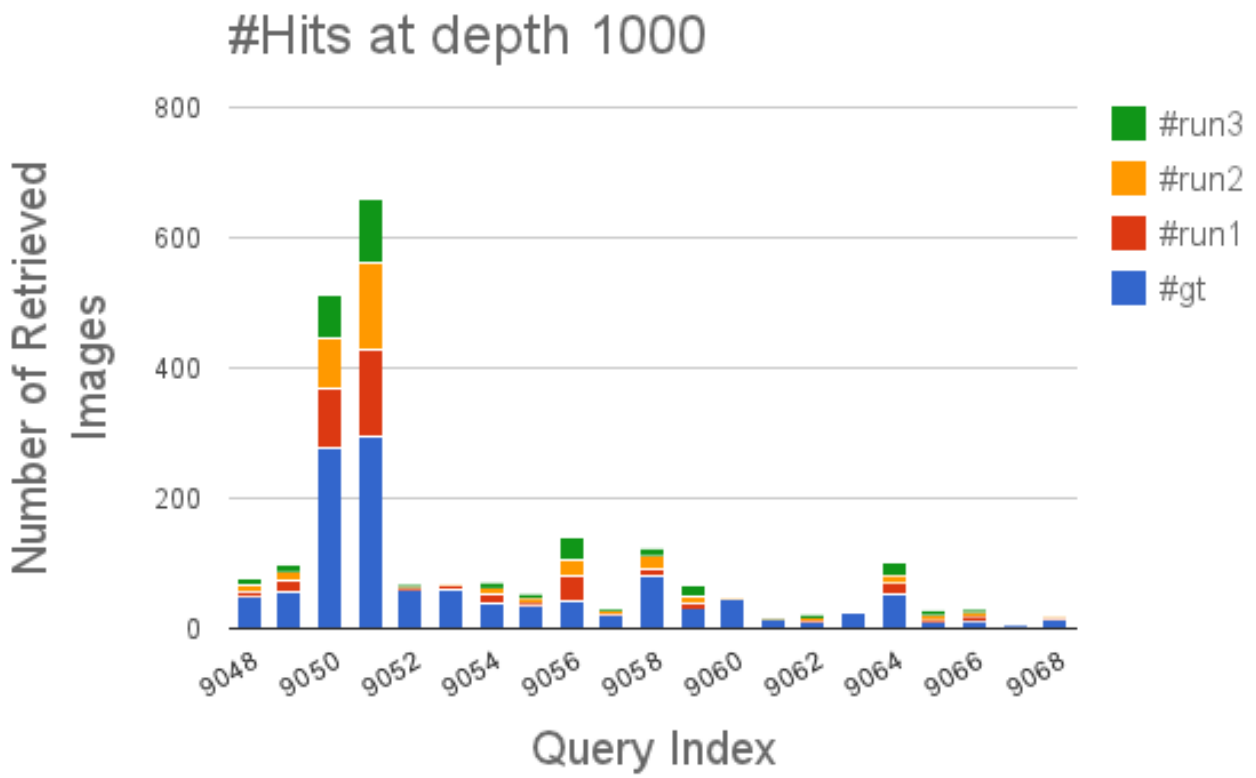

Fig. 4. Number of hits at the depth of 1000 images for Run-1, Run-2 and Run-3 compared to the ground truth (gt). Best in color.

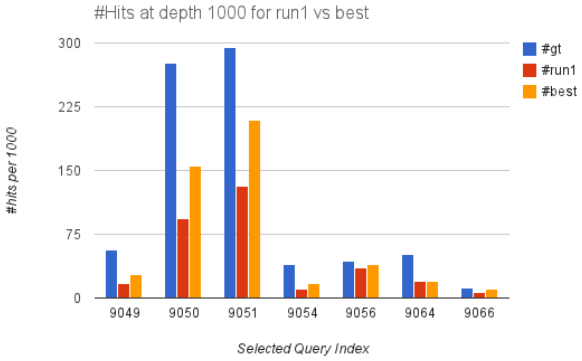

Fig. 5. Number of hits at the depth of 1000 images for Run-1 compared to the ground truth (gt) and the best for the selected query topic

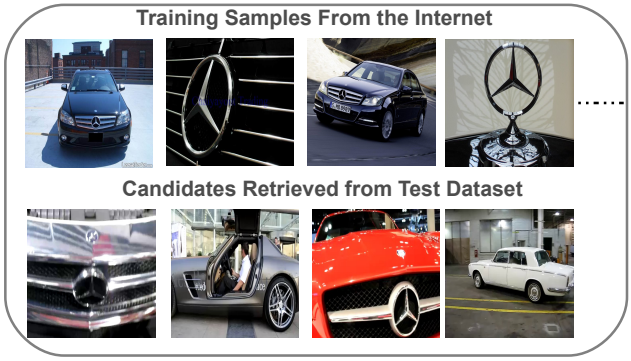

Fig. 7. Top row shows some of the training images crawled from for learning the discriminative classifier. Bottom row shows some of the retrieved images at a depth of 1000. As seen, the contextual information plays a major role in improving the accuracy.

learning the discriminative model. Figure 7 illustrates how the context information is helpful in improving the retrieval using discriminative re-ranking.

*C. **Run 3**: BoW with Histogram Intersection + SVM trained on PHOW*

For the third set of experiments, we used the Bag of visual words model with Chi-square distance for matching the given query images with the key-frames in the database and re-ranked using the classification scores obtained by SVM classifier trained on internet images with color PHOW features. Results obtained from run-3 are similar to that of run-2 since the discriminative model learned from the internet images did not adapt well in the test domain.

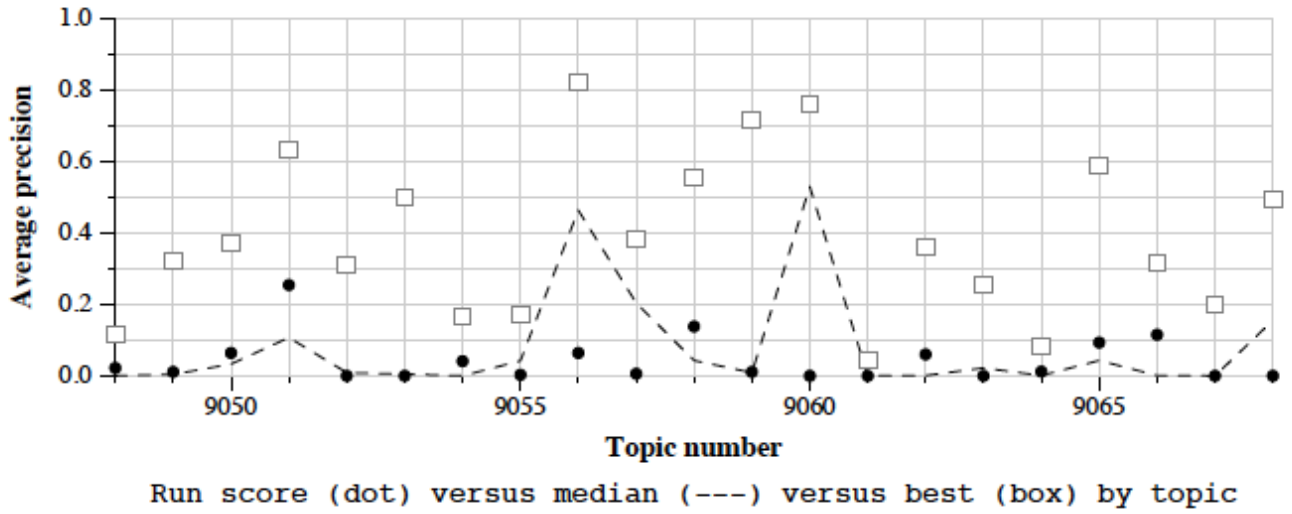

Fig. 6. Shows mean average precision obtained for different topics with Run-2.

V. Conclusion

For the instance search task, a bag of visual words model (BOW) based retrieval strategy is effectively coupled with discriminative linear classifiers for re-ranking. The training dataset from 2011 instance search task is used for learning the dictionary and the learned dictionary is used on the test dataset for obtaining the image descriptors. Because of the missing labels in the test dataset, we crawled additional training images from Google and FLICKR image databases using textual meta-data available along with query images to train a linear classifier using Support Vector Classifier (SVM). The classification margin is used for scoring the query class likelihood for every key-frame image sampled from the video shots.

Initial list of candidate images are retrieved using BOW model and chi-square distance metric, and then SVM classifier learned from the internet images is used to re-rank the candidates. In order to reduce the overall retrieval time, linear classifiers are run against the test database in an offline manner. Experimental results show better performance for global queries that occupy considerable portion of the image plane. Also, due to inherent differences in the data distributions of the training and test datasets, the discriminative model learned from web images did not perform as good as it performed on the training source domain. In the future, we plan to perform further research on Domain adaptation i.e. how to transfer models from the source domain to the target domain automatically.

Acknowledgment

The authors would like to thank Yue Cao and Michael Shabsin for the help during the course of the task.

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918. IEEE, 2012.

[2] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[3] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465. IEEE, 2005.

[4] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[6] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[8] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

[9] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[10] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012.