Exploring The Optimal Visual Vocabulary Sizes for Semantic Concept Detection

Jinlin Guo, Zhengwei Qiu, Cathal Gurrin
CLARITY and School of Computing
Dublin City University, Dublin, Ireland

***Abstract*—The framework based on the Bag-of-Visual-Words (BoVW) feature representation and SVM classification is popularly used for generic content-based concept detection or visual categorization. However, visual vocabulary (VV) size, one important factor in this framework, is always chosen differently and arbitrarily in previous work. In this paper, we focus on investigating the optimal VV sizes depending on other components of this framework which also govern the performance. This is useful as a default VV size for reducing the computation cost. By unsupervised clustering, a series of VVs covering wide size range are evaluated under two popular local features, three assignment modes, and four kernels on two different-scale benchmarking datasets respectively. These factors are also evaluated. Experimental results show that best VV sizes vary as these factors change. However, the concept detection performance usually improves as the VV size increases initially, and then gains less, or even deteriorates if larger VVs are used since overfitting happens. Overall, VVs with sizes ranging from 1024 to 4096 achieve best performance with higher probability when compared with other-size VVs. With regard to the other factors, experimental results show that the OpponentSIFT descriptor outperforms the SURF feature, and soft assignment mode yields better performance than binary and hard assignment. In addition, generalized RBF kernels such as χ^2 and Laplace RBF kernels are more appropriate for semantic concept detection with SVM classification.**

I. Introduction

Currently, the emphasis on detecting semantic concepts now is moving to more generalized semantic indexing. In particulary, a recent trend in semantic concepts detection has been to search for generic methods that are based on BoVW feature representation [1] and Support Vector Machine (SVM) framework, which has produced significant results on several large scale content-based image and video retrieval benchmarkings, such as TRECVid [1]. The BoVW model allows semantic concept detection by representing an image by a distribution of visual words (VWs) defined beforehand in a VV. Fig. 1 shows the generation process of BoVW-based feature representation and SVM classification. The BoVW model is inspired by the bag-of-words approach to text-document categorization. However, compared with textual-document categorization, there is no available vocabulary for image-based semantic concept detection and it has to be learned from a training image set.

[1]In the literature, the *bag of feature* model, the *codebook* model and the *BoVW* model are nearly the same. The terms " textons","visual words" and "codewords" have been used with approximately the same meaning, i.e. clusters of feature space in a high-dimensional space, although "textons" is usually used in texture recognition. In this paper, we use visual words and visual vocabulary representing the two cores of this model

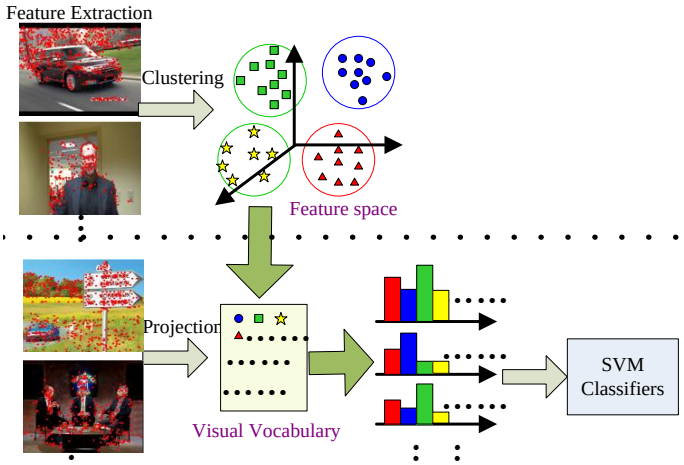

Fig. 1. Framework based BoVW feature representation and SVM classification for semantic concept detection

TABLE I. An overview of different VV sizes used in previous literature

	Size		Size
[6]	6000∼10000	[7]	300,1000,4000
[8]	200∼400	[9]	1000
[10]	< 1500	[2]	500∼2500

The VVs are generally constructed by using unsupervised algorithms to quantize the high-dimension low-level feature descriptors into clusters [2], [3], among which, *K*-means is the most popularly used because of its simple and easy implementation. However, the number of clusters *K*, in unsupervised clustering algorithms is typically chosen arbitrarily, which makes the VV sizes used in previous work always different, and there is no consensus which size or range of VV sizes can produce the best performance.

About this framework, there has already been a lot of existing reviews, which can be found in [3], [4], [5]. Here, we only give a brief review of the literature about the choice of VV size in previous work. An overview of different VV sizes found in literature is in Table I. In [3] Jiang et al. analyzed each aspect of the BoVW model in depth, and experimented with VVs of various sizes on TRECVid2006 datasets under different choice of weighting schemes (assignment modes) and kernel choices for SVM. They found that appropriate size of the VV for different datasets are different. Csurka et al. in [2] reported that performance improves steadily as the VV size grows range from 500 to 2500 using the keypoint features (SIFT), which

were also found in [7] not only using SIFT but also considering dense sampling. However, as the VV size is increased, apparent overfitting was found in [7]. Both Li et al. [10] and Gemert et al. [5] aimed to construct compact and optimal VVs for object recognition or visual categorization. They also experimented with different VV sizes under different factor settings. In [10], the experimental results show substantial gains in performance as the VV size increased. Gemert [5] obtained similar results but with larger range of VV size (with a maximum size of 3223). In [4], Chatfield et al. also evaluated this framework with different settings. The number of visual words they used varies between 600 and 8,000 (specifically, 600, 1500, 2000, 4000, and 8000) for the Caltech-101 data and between 4,000 and 25,000 for the PASCAL VOC data. Their experimental results demonstrate that larger vocabularies lead to higher accuracy.

However, a large number of VWs also require extra processing overhead, such as computation storage and time. Specifically, it takes more time to produce a larger VV, then consequently it will take more time to map the low-level feature descriptors to VWs, and to train a classification model. Therefore, it is important to choose optimal VV size(s) for efficient concept detection. In this paper, we evaluate the detection performance of a series of VVs with different sizes under other factors including the choices of local feature, assignment mode of VWs, and kernel function for SVM learning on two datasets. Compared with previous work, our contributions in this work are that we 1) focus on find the optimal ranges of VVsize for concept detection across different factors, i.e. datasets, local feature, assignment mode and kernel selection. Extensive experiments are performed on both video datasets and image datasets, and the impact of these factors are also evaluated; 2) evaluate the performance of a large range of VV sizes, which nearly covers all the VV sizes having been used in previous literature.

In the following, section 2 outlines the details about constructing the representation of BoVW and SVM classification for the experiments in section 4. Experimental details and result analysis are described in Section 3 and 4 respectively. Finally, Section 5 concludes this paper.

II. Factors in BoVW model and SVM Framework for Semantic Concept Detection

In the BoVW and SVM framework for concept detection, the performance varies as different-size VV is used. Furthermore, choices of local feature, assignment mode and kernel also govern the performance. Our experiments in this paper try to investigate the concept detection performance of VVs with various sizes under different local features, assign modes and kernels. Hence in this section, we analyze these four factors and especially describe our choices that will be evaluated in later experiments.

A. Local Features

The major problem for automatic concept detection is bridging the *semantic gap* between low-level feature representations extracted from sensor data (images and videos) and high-level human semantic interpretation of the data. Hence, visual features need to model the wide diversity in appearance of semantic concepts. There are also variations in appearance which are not due to the richness of the semantics. Varying the viewpoint, lighting changes, clutter and occlusions in the recording of a scene will deliver different data, whereas the semantics may not have changed. Recently, there is a trend of using image scale- or affne-invariant local feature keypoints, which are proved effective for semantic concept detection by consecutive-reported progresses [3], [11]. These invariant visual features are computable visual properties that are insensitive to changes in the content, for example, caused by changing the color illumination intensity, rotation, scale, translation, or viewpoint, while still able to distinguish concepts with different semantics.

Different feature detectors, nevertheless, emphasize different aspects of invariance, resulting in keypoints of varying properties. Here, we evaluate two popular keypoint descriptors which also produced good performance in previous work, including OpponentSIFT (OppSIFT) [11] descriptor and SURF [12] feature.

B. VV Construction and Sizes

VVs are popularly constructed by the simple square-error based partitioning method: *K*-means. However, there are some deficiencies. For example, one is that it easily converges only to local optima. Another is that it does not determine the parameter *K*. Moreover, two practical shortcomings are it is computationally expensive and the memory required for *K*-means would be prohibitive for VV and training sets of large scale.

In this paper, we construct a large series of VVs which cover a wide range of sizes. Hierarchal *K*-means technique inspired by [13] is used to construct the VV tree. Instead of *K* defining the final number of clusters or quantization cells, *K* define the branch factor (number of children of each node) of the tree. Firstly, an initial *K*-means process is run on the training data, defining *K* clusters. The training data is then partitioned into *K* groups, where each group consists of the feature vectors closest to a particular cluster center. Then the same process is then recursively applied to each group of feature vector, recursively defining quantization cells by splitting each quantization cell into *K* new parts. The VV tree is determined level by level, up to some certain number of levels *L* and each division into *K* parts is only defined by the distribution of the feature vectors that belong to the parent quantization cell. Advantages of this construction method are mainly as follows: 1) It speeds up the assignment stage and 2) It is easy to construct large-size VVs with limited computational resources (such as RAM). When assigning, each feature vector is simply passed down the tree by at each level comparing the feature vector the the *K* candidate cluster centers (represented by *K* children in the tree) at the i_{th} level and choosing the closest one. This is a simple matter of performing *K* similarity computations at the level, resulting in a total of $L*K$ similarity computations; Furthermore, based on our internal experiments, the performance differences are very small when comparing VVs constructed by this method with simple *K*-means.

Based on the operability and practicality, we construct VVs with the sizes illustrated in Table II. Choosing these sizes also

TABLE II. VV OF DIFFERENT SIZES CONSTRUCTED BY *K* AND *L* IN THIS PAPER BY HIERARCHAL *K*-MEANS

L, *K*	4	5	6	7	8
3	64	125	216	343	512
4	256	625	1296	2401	4096
5	1024	3125	7776	16807	32768
6	-	15625	-	-	-

involves the trade-off among description ability, computational efficiency and coverage of VV sizes used in previous work. In practice, for each datasets and each low-level feature, we sample the training sets and cluster 2,000,000 feature vectors respectively. Furthermore, Hierarchal K-means clustering may introduce additional discretization error. In order to overcome this defect, we run it 20 times with different initial centers, and select the one with the least variance.

C. Assignment Modes

Most of the existing work adopts the nearest neighbor search in the VV in the sense that each feature descriptor is assigned to the most similar VW(s). Here, we investigate the performance of VV with various sizes under three popular assignment modes: binary, hard and soft assignment. Binary assignment is indicating the presence and absence of a VW with values 1 and 0 respectively. Hard assignment is simply counting the presence of the VWs. Soft assignment was reported to achieve better performance in [3], [5]. In our experiment, we employ the soft assignment method proposed in [3], which was reported to perform well for concept detection. For each keypoint extracted in an image, instead of assigning it only to its nearest visual word, in soft assignment the top-4 nearest visual words are selected. Suppose we have a VV of *K* VWs, we use a *K*-dimensional vector $[\omega_1, \omega_2, ...\omega_K]$ with each component representing the weight ω_t of a visual word *t* in an image such that

$$\omega_t = \sum_{i=1}^{4} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim\left(j, t\right) \tag{1}$$

where M_i represents the number of keypoints whose i_{th} nearest neighbor is the VW *t*. The measure $sim(j,t)$ represents the Cosine similarity between feature descriptor *j* and the VW *t*. In this equation, the contribution of a feature descriptor is its similarity to VW *k* weighted by $\frac{1}{2^{i-1}}$ representing that the VW is its i_{th} nearest neighbour.

D. Kernels for SVM Classification

The learning ability of a SVM classifier depends on the type of kernel used. In this paper, we investigate the performance of VVs with various sizes using different kernels for SVMs, including Linear kernel, traditional Gaussian Radius Basis Function (RBF) kernel and two generalized RBF kernels because of either their efficiency or their good performance in concept detection using SVMs.

- Linear Kernel:

$$K\left(\boldsymbol{x}, \boldsymbol{y}\right) = \boldsymbol{x}^T \boldsymbol{y} \tag{2}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two input vectors.

- Generalized forms of RBF kernels:

$$K\left(\boldsymbol{x}, \boldsymbol{y}\right) = e^{-\gamma d(\boldsymbol{x}, \boldsymbol{y})} \tag{3}$$

where $d\left(\boldsymbol{x}, \boldsymbol{y}\right)$ can be chosen to be any distance measurement in the feature space. Since BoVW representation is a histogram of visual words with discrete densities, the χ^2 distance may be more appropriate:

$$d_{\chi^2}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{1}{2} \sum_i \frac{\left(x_i - y_i\right)^2}{x_i + y_i} \tag{4}$$

which gives a χ^2 RBF kernel.

In addition to χ^2 kernel, there are other generalized RBF kernels with the distance function defined as:

$$d_b\left(\boldsymbol{x}, \boldsymbol{y}\right) = \sum_i \left|x_i - y_i\right|^b \tag{5}$$

With this distance function, equation becomes the Laplacian RBF kernel when $b = 1$, and the traditional Gaussian RBF kernel while $b = 2$.

III. EXPERIMENTAL SETUP

The experiments focus on investigating the concept detection performance of VVs with various sizes under different visual local features, assignment modes and kernels across different datasets, which aims to find the best size or range of VV size. These factors are also evaluated. To obtain a persuasive conclusion, we identify three experiments according to considered factors:

1. Local Image feature

2. Assignment mode

3. Kernel choice

These experiments are conducted on two benchmarkings datasets. Ground-truth is is obtained either by manually annotating if a concept is present in the images (or video shots) or by pooling. This fixed ground-truth allows repeatable experiments.

To reduce dependency on datasets, we evaluate VVs of various sizes on two different-scale datasets, MIRFLICKR [14] and TRECVid 2010 [1]. The MIRFLICKR is an image benchmarking while TRECVid is a video benchmarking exercise. For the image benchmarking, We choose the MIRFLICKR-25000 (MR) set. VVs of various sizes are evaluated on this dataset using five-fold cross validation on the 25,000 images. At each fold, SVM classifiers for each concept are trained on 15,000 of the data, and then tested on the remaining 10,000, selected at random. Furthermore, real ground truth is available for the entire set from [14].

The TRECVid 2010 datasets (TV) consists of 264,615 keyframe images extracted from about 100 GB video data. 119,685 images and 144,931 images are used for training set and test set respectively. Since the annotations for the entire dataset are not provided, here, we only train SVM classifiers on the provided training set and evaluate the performance on test set, that is, no cross validation is performed on this dataset.

These concepts detected in MR datasets and TV datasets are listed in Table III. These concepts are selected since

TABLE III. CONCEPTS DETECTED IN MR DATASETS AND TV DATASETS RESPECTIVELY

	Concepts
MR	baby, bird, car, clouds, dog, female, river, sea, flower, male, night, people, portrait, tree
TV	animal, boat/ship, explosion fire, mountain, vehicle, swimming, bus, airplane flying, car racing, dancing, sitting down, hand, doorway, running, cheering, Asian people, cityscape, classroom, old people, walking, ground vehicle, demonstration/protest, female human face closeup, nighttime, telephone, throwing, dark skinned people, flowers, bicycling,

we only have their ground truth or sampled truth data. For evaluation, we use the average precision (AP) for MR datasets and common measure inferred average precision (infAP) for the TV datasets [15]. To aggregate the performance of multiple semantic concepts, mean AP (MAP) and mean infAP (MinfAP) are used for MR and TV datasets respectively.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we report and analyze the results from three experiments using different size of VVs under different local features, assignment modes and kernels respectively.

A. Experiment 1: Local Image Feature

The first experiment compares concept detection performance of different-size VVs using two local features. Our experiments show that χ^2 kernel with soft assignment always produces better performance than Linear kernel and Gaussian kernel, and comparable performance with Laplace kernel. Here we only report results when using χ^2 kernel and soft assignment on two datasets.

Experimental results are shown in Fig. 2. The figure illustrates that the performance changes a lot when different-sized VVs are employed. For example, in TV datasets, when using OppSIFT descriptors, VV with the size of 1296 increases the performance by 70.1% comparing with VV of 64 VWs, whereas VV with 625 outperforms by 40.5% than VV with 64 VWs in MR dataset. It also shows that the performance improves as VV size increases overall for both the features and datasets. The best VV sizes vary as features and datasets. However, on both datasets, results show that if using larger VVs than a certain size gains little or even deteriorates the performance since overfitting happens. More specifically, for MR datasets, on both features, it achieves best MAPs when VV sizes range from 512 to 4096. The MinfAPs increase steadily when VV size increases from 64 to 4096 for TV datasets for both features, and it achieves best when VV sizes are between 512 and 7776.

Over two different-scale datasets, Fig. 2. show that the OppSIFT achieves better performance for all VV sizes than SURF, which proves that SIFT-like, OpponentSIFT can produce better performance than SURF. Furthermore, much better performance is reported when using MR datasets than TV datasets for all the VVs which are also illustrated in later experimental results. We speculate this is because: 1) the annotation quality of MR training datasets is much better than TV datasets. The MR images were annotated carefully by experts, while TV training datasets are much larger and

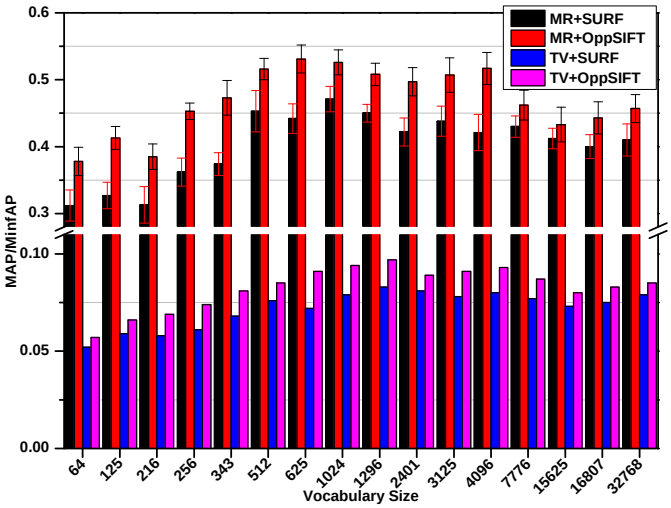

Fig. 2. Performance of VVs with different sizes using two local features on two datasets. Bars and error-bars indicate the mean and standard deviation over five-fold cross-validation testing respectively.

were collaboratively annotated by a number of different teams. Previous work [16] have shown that annotation quality significantly affects the performance; 2) the MR datasets are picked up from photos uploaded by thousands of individuals, and represent the image retrieval area much more effectively; 3) there are several action concepts in the concept set of TV, however, in this work we only use static features which cannot capture the space-time aspects that are more effective to detect the action; 4) another important reason is that concepts for TV are much more infrequent than those of MR; 5) the most important but not the last reason is the two evaluation modes are different. MR evaluates the results according to the ground truth for the entire test set, while TV use ground truth obtained though pooling, only images are annotated that appear in the top N (2000) of most relevant images in the ranking of at least one approach participating in the benchmarking. Pooling reduces the cost of annotation considerably, but it may leave large parts of collections unlabeled, thereby hindering accurate measurement of precision.

The differences per concept among different-size VVs are shown in Fig. 3. Here we only show the four top-performance concepts for OppSIFT features on two datasets respectively. As shown in Fig. 3, for different concepts, best-size VVs are different. On the whole, the infAPs/APs improves consistently with MinfAPs/MAPs as VV size increases. Specifically the performance increases initially as the VV become larger, then gains less or even worsens if VV continuously increases. Based the variations of performance, VVs with sizes from 343 to 7776 achieve optimal with high probability.

B. Experiment 2: Assignment Modes

In this experiment, we focus on the impact of VV size on detection performance when different assignment modes are adopted. Here we report the results by using OppSIFT feature and χ^2 kernel for SVM on MR datasets. Results from SURF feature and the other kernels and TV datasets are similar. In Fig. 4, we show the results, with a similar observation as shown in Experiment 1. Increasing the number of VWs

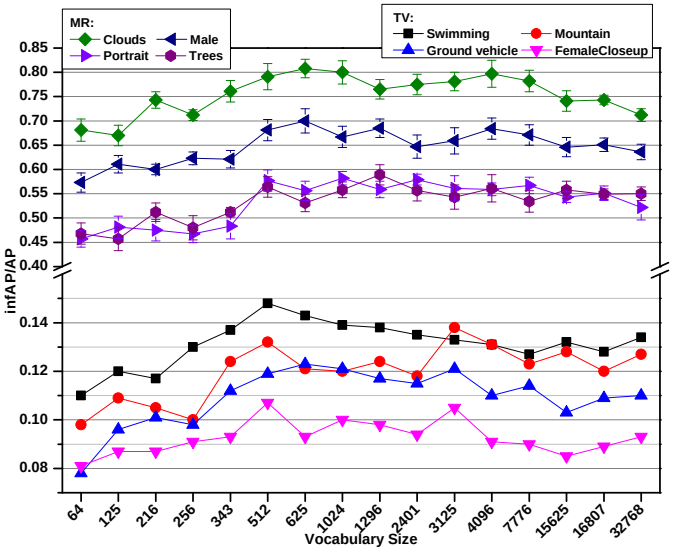

Fig. 3. Performance of the top concepts across two datasets using OppSIFT features in Experiment 1

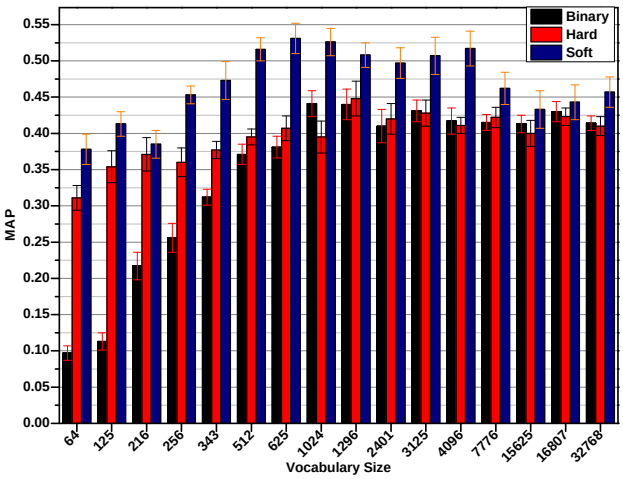

Fig. 4. Performance of VVs with different sizes using three assignment modes in MR datasets

firstly promotes the performance for three assignment modes. Then it gains a little or even degrades the performance and the performance gain at a price of paying more computational resources and time.

Next, let us examine the impact of different VV sizes in detail. When using binary assignment, we observe that the best VV sizes range from 1024 to 4096. For hard assignment, the performance fluctuates as different-size VVs are adopted, and VV sizes in the range of 343 and 7776 are the appropriate compromise. While using soft assignment, the performance is higher when VV sizes are from 256 to 7776.

Concerning the performance comparisons between binary, hard and soft assignment modes and different-size VVs, we find that initially the performance increases drastically when larger VVs as well as binary assignment are used. Hard assignment outperforms binary assignment by a large margin only when the VV size is small, which due to the fact that, with a larger VV size, the count of most VWs is either 0 or 1, and thus similar with binary assignment. Across all VV sizes and datasets, soft assignment outperforms the other two assignment modes. We speculate that assigning a keypoint only to its nearest neighbor VW may be not an optimal choice, given the fact that two similar keypoints may be clustered into two different clusters when increasing the size of the VV, that is with slight variation in the images, hard assignment may choose complete different VWs. This may also explained by such an example, instead of labeling a blue local patch as *sky*, the patch is better represented by saying that its similarity to *sky* is 0.9, and its similarity to *water* is 0.8. Therefore, soft assignment is robust, which are proved by the experimental results shown in Fig. 4. When the VV sizes are in the best range, the MAPs of the soft assignment vary just in a smaller range than binary and hard assignment.

C. Experiment 3: Kernel for SVM

In this experiment, we move on to investigating the impact of different kernels in SVM on concept detection performance accross different VV sizes. Here, we only report the results of using OppSIFT feature and soft assignment on TV datasets. Fig. 5 summarizes the performances of various kernels accross all VV sizes and datasets. The results of SURF feature and other assignment and TV datasets are similar. For the generalized RBF kernels, we vary the parameter γ in a range from 2^{-7} to 2^{3} and choose the best one via 5-fold cross validation. As shown in Fig. 5, we can get similar observation as Experiment 1 and 2. We observe that an appropriate range of VV size is from 256 to 15625 for different kernels and datasets.

Overall, the Gaussian RBF kernel and generalized RBF kernels perform much better than the Linear kernel. This indicates that the concept classes are correlated to each other in BoVW feature space and thus are not simply linearly separable. However, when using the Linear kernel, the performance improves marginally as the VV size increases initially. Then it nearly keeps the same as VV size continually increases. This shows that a limited gain can be achieved by Linear kernel even the BoVW features are not linearly separable.

More interesting observations are from the performance of Gaussian and the other two generalized RBF kernels. The results show that the three RBF kernels achieve comparable performance. On the whole, the generalized RBF kernel χ^2 and Laplace kernel consistently outperform the traditional Gaussian RBF kernel. In the case of χ^2 and Laplace kernel, they used to achieve more or less the same performance. We deem that this is because these two kernels are both linear exponential decay and tolerate the background variance without amplifying the effect, while emphasize the regions only containing the target concept. In addition, with regard to efficiency, the computation time for Linear kernel is the fastest since no exponential computation is needed.

D. Summary of Experimental Results

The first observation we can make is that the best VVs are different for different local features, assignment modes and kernels across different datasets for semantic concept detection. However, the performance improves initially as the VV size increases, then gains a little or even worsens if VV

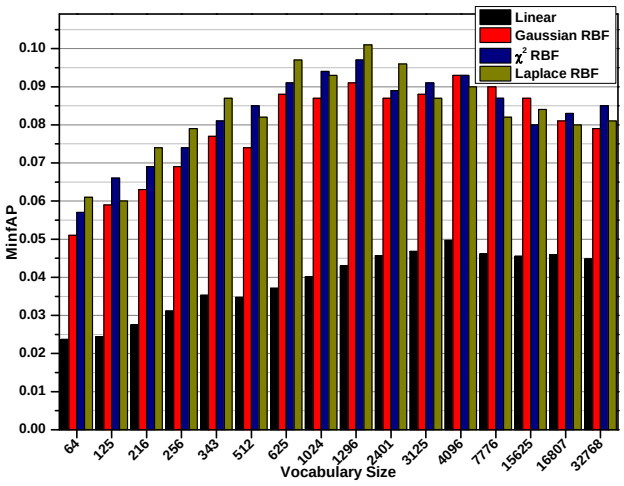

Fig. 5. Performance of VVs with different sizes using four kernels for TV dataset

TABLE IV. SUMMARY OF THREE EXPERIMENTS WITH DIFFERENT FACTORS TO OBTAIN THE BEST RANGE OF VV SIZE.

Factor selection	Best-size Range
MR+SURF/OppSIFT+Soft+χ^2	512∼4096
TV+SURF/OppSIFT+Soft+χ^2	512∼7776
MR/TV+OppSIFT+Binary+χ^2	1024∼4096
MR/TV+OppSIFT+Hard+χ^2	343∼7776
MR/TV+OppSIFT+Soft+χ^2	256∼7776
MR/TV+OppSIFT+Soft+Gaussian	625∼15625
MR/TV+OppSIFT+Soft+Laplace	343∼4096

continuously increases since overfitting is found. In Table IV, we summarize the best ranges of VV size in above experiments roughly based on the observations of experimental results. As shown in Table IV, for both local features, the VVs with the sizes range from 512 to 4096 may report better performance. In term of three assignment modes, we suggest the VV sizes from 1024 to 4096 as a safe application. VV sizes among 1024 to 4096 are better for three different kernels. Overall, we conclude that VVs with sizes ranging from 1024 to 4096 typically achieve best performance with higher probability comparing with other-size VVs.

For these factors local features, assignment modes and kernels which also dominate the performance, experimental results show that SIFT-like feature descriptor OppSIFT outperforms SURF feature, and soft assignment mode yields better performance than binary and hard assignment and generalized RBF kernels such as χ^2 and Laplace RBF kernels are more appropriate for semantic concept detection with SVM classification.

V. CONCLUSIONS

In this paper, we investigate and evaluate the semantic detection performance of a series of VVs with different sizes by jointly considering factors local feature, assignment mode and kernel. Meanwhile, these factors are also evaluated. We experimentally show that best VV sizes vary as different local feature, assignment mode and kernel are used. However, the performance usually improves as the VV size increases initially, then gains a little even deteriorates if larger VVs are used. On the whole, VVs with sizes ranging from 1024 to 4096 achieve best performance with higher probability comparing with other-size VVs. Our experimental results also show that factors such as local feature, assignment mode or kernel are influential to the performance. Specifically, local feature descriptor OppSIFT outperforms SURF feature, and soft assignment mode yields better performance than binary and hard assignment and generalized RBF kernels such as χ^2 and Laplace RBF kernels are more appropriate for semantic concept detection with SVM classification.

ACKNOWLEDGEMENTS

Thanks to the Information Access Disruptions (iAD) Project (Norwegian Research Council), Science Foundation Ireland under grant 07/CE/I1147 and the China Scholarship Council for funding.

REFERENCES

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. of MIR*, 2006, pp. 321–330.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and E. Bray, "Visual categorization with bags of keypoints," in *Proc. of Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[3] Y.-G. Jiang, J. Y. 0003, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[4] A. V. Ken Chatfield, Victor Lempitsky and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. of the BMVC*, 2011, pp. 76.1–76.12.

[5] J. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Comparing compact codebooks for visual categorization," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 450–462, 2010.

[6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003, pp. 1470–1477.

[7] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. of ECCV*, 2006, pp. 490–503.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of CVPR*, 2006, pp. 2169–2178.

[9] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[10] T. Li, T. Mei, and I.-S. Kweon, "Learning optimal compact codebook for efficient object categorization," in *Proc. of WACV*, 2008, pp. 1–6.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, 2010.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[13] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of CVPR*, 2006, pp. 2161–2168.

[14] M. J. Huiskes and M. S. Lew, "The MIR FLICKR retrieval evaluation," in *Proc. of MIR*, 2008, pp. 39–43.

[15] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proc. of CIKM*, 2006, pp. 102–111.

[16] S. Nowak and M. J. Huiskes, "New strategies for image annotation: Overview of the photo annotation task at ImageCLEF 2010," in *Proc. of CLEF*, 2010.