MESH participation to TRECVID2008 HLFE

J. Molina[3], V. Mezaris[2], P. Villegas[1], G. Tolias[4], E. Spyrou[4], N. Sofou[4], J. Rodríguez-Benito[1], G. Papadopoulos[2], S. Nikolopoulos[2], J. M. Martínez[3], I. Kompatsiaris[2], P. Kapsalas[4], A. Dimou[2], E. Bru[1], Y. Avrithis[4], T. Adamek[1]

[1]Telefónica I+D, Spain

[2]Informatics and Telematics Institute/Centre for Research and Technology, Greece

[3]Universidad Autónoma de Madrid, Spain

[4]National Technical University of Athens, Greece

Abstract

A group of four organizations from the MESH consortium (**www.mesh-ip.eu**) participated this year for the first time in the High Level Feature Extraction track in TRECVID. The partners were Telefónica I+D (TID, Spain), Informatics & Telematics Institute (ITI, Greece), National Technical University of Athens (NTUA, Greece) and Universidad Autónoma de Madrid (UAM, Spain). We submitted a total of 6 runs, using different variations and configurations over a common model.

With only one exception, results obtained by those runs were below expectations, mostly due (we believe) to some implementation bugs discovered afterwards. Some of those errors have already been solved and we hope to correct the rest and improve the performance of the system for future editions.

1. Introduction

This is the first participation of the partners in the MESH consortium in the TRECVID HLFE track (though some of them had previous experience in past editions separately). The MESH project developed a common visual analysis infrastructure to detect high level concepts in visual scenes; though the set of concepts was only partially coincident with those of TRECVID 2008 (in MESH it is tuned to news content). The system had then to be adapted and trained for the concept set in TRECVID, and for the MAP metric used for evaluation here. In the course of the development a few new techniques, not originally present in the MESH system, were also tried.

With only one exception (a motion activity computed over the video stream) all the remaining data extracted from the media was done on still keyframes; for those the reference shot segmentation provided by Fraunhofer-HHI for TRECVID 6 was used. We did not use audio information for any of the runs.

The main architecture of the HLFE system is based on well-known paradigms in visual analysis, such as MPEG-7 descriptors, SIFT interest points and SVM classifiers. Nevertheless, we hoped that the specifics of their combination would provide good results. Moreover, one guiding principle in the development was not to include human intervention in model selection and configuration for each individual feature. This rule stems from our aim to be able to generalize the system to any additional feature without resorting to human intelligence to select and combine adequately the available set of tools. The system, thus, gets trained blindly with a ground-truth training set, and adapts automatically to the specifics of each concept during this training phase.

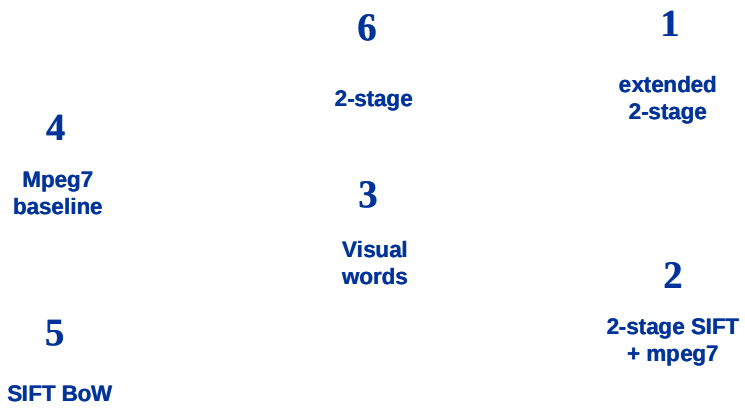

Figure 1 : run configuration and dependencies

2. Overall structure of the submission

As commented, a total of 6 runs were submitted for evaluation. All of them share a base common infrastructure, which uses SVM classifiers (provided by LIBSVM 6) fed with different configurations of features extracted from keyframes and video. Training was done using the ground truth provided by a collaborative annotation effort 6.

The set of runs was scheduled so that most of them are interrelated, re-using part of the results and systems of previous ones, plus a couple of standalone runs developed separately (though still using that same architecture of SVM classifiers). The dependency scheme is shown in figure 1; in what follows each run is described briefly (ordered by complexity and dependency):

- The *MPEG7 Baseline run* (run 4), produced by TID is intended to provide the baseline, by means of a very simple approach with binary SVMs trained with MPEG-7 descriptors and a majority decision.
- The *SIFT BoW run* (run 5) was done by TID using random SIFT descriptors with the “Bag of Words” approach as input for the classifiers.
- The *Visual Words run* (run 3) is the first one (run 4, baseline) with 6 concepts changed using an approach based on Visual Words developed by NTUA (concepts numbered 1, 6, 7, 10, 16, 17 of those defined in TRECVID 2008).
- The *2-stage run* (run 6) was developed by UAM over the Baseline run by adding a second step of SVM classifiers.
- The *Extended 2-stage run* (run 1) was done by TID & UAM by adding additional descriptors (motion, faces, persons) into the structure of the previous run.
- The *2-stage SIFT+mpeg7 run* (run 2), made by ITI, is based on a combination of MPEG-7 and SIFT-based 6 global image features.

The following section provides more details about each of these runs, followed by a brief account of the initial analysis of the results obtained with each one.

3. Description of runs

3.1. MPEG-7 baseline (run 4)

Run 4 is a baseline classifier that uses global MPEG-7 visual descriptors as separation patterns. The following MPEG-7 descriptors were used: color layout, edge histogram, homogeneous texture and scalable color. The training process

was performed using the annotated keyframes provided through the collaborative annotation effort.

Normalization is performed by using the covariance matrix of the feature vector (pre-computed with all feature vectors extracted over a large content test set). This equalizes the dynamic range of all descriptors. It also helps to reduce the statistical (linear) dependency between the different descriptors, and thus improve the discriminative power of the feature vector.

Support Vector Machines are used as classifiers. Libsvm 6 was used as the SVM implementation, with a Radial Basis Function Gaussian kernel.

An SVM classifier was trained to produce classification models for every High Level Feature. Training is performed in a fully automatic way, and it involves a dynamic selection of

- The optimal subset of the content set that will be used for training, probing with different combinations of the ratio positive/negative examples
- The combination of MPEG-7 descriptors to include in the input vector, using full tests over a set of 6 possible configurations.
- The SVM parameters C, γ, using a logarithmic grid in the parameter space (akin to the tool also provided by LIBSVM)

Selection of the best combination is done through 5-fold cross validation over the entire training set, optimizing Mean Average Precision (though, as commented later, there was a bug in this optimization phase for the submitted run). The statistical validation allows the SVM to output a confidence level when working on the test set.

This first run used simple threshold setting on the SVM output for deciding the presence of a High Level Feature in a frame.

3.2. SIFT Bag of Words (run 5)

Run 5 is based on SIFT descriptors 6 and the “Bag of Words” (BoW) methodology 6. 10000 points are randomly selected within every image 6 and the area around every point is described using SIFT descriptors. A Visual Word Vocabulary (Codebook) is created using hierarchical K-Means clustering of SIFT points from the development set resulting in a tree with 10000 leafs 6. The tree allows representation of images as BoW by rapid mapping of key-points to visual words from the vocabulary. Finally, images are represented as binary histograms containing information about presence or absence of visual words 6.

A set of SVM classifiers are then trained on the local features based on those SIFT descriptors and the BoW approach, using LIBSVM 6. The SVM parameters are optimized without supervision by maximizing Mean Average Precision over the training set.

3.3. Visual Words (run 3)

Run 3 is based on a part-based object detection method 6 and a visual words Approach 6. The first method is used to detect *Airplane_flying* and *Two People,* while the second is used to detect *Classroom*, *Cityscape*, *Mountain*, and *Nighttime*. This run uses supplemental training data in addition to NIST development data and annotations (run type C).

The part-based object detection method is based on Haar-like and Harris features. The detectors are trained to recognize parts of objects. For example, for the detection of a Person, a face, an upper and a lower body part detector. To

Figure 2 : Face detection module for the Extended 2-stage run

train the detectors for this method, the PASCAL dataset 6 been used, along with a part of the TRECVID 2008 development data.

The visual words approach begins with an initial RSST color segmentation algorithm which produces a coarsely segmented image. A visual dictionary is constructed by clustering the extracted regions based on their MPEG-7 color and texture features. Using this dictionary, a given is then described in terms of the produced visual words. The training of the detectors has been performed on the TRECVID 2008 development data. An SVM detector has been trained for each concept. In an effort to improve the quality of the results, several contextual relations 6 such as co-occurrence have been applied.

The detection of the 14 remaining high-level concepts of TRECVID in this run was performed in the same way it was done in the baseline run.

3.4. 2-stage classification (run 6)

Run 6 is based on the MPEG-7 Baseline. An extra SVM is used per concept, using as input not only the previous prediction for that concept, but the degree of confidence of the baseline predictions for all the twenty concepts. This way, the correlation within the concepts under consideration is taken into account for improving individual classification results.. For example, a high degree of truth on positively detecting two persons might be important to detect whether the shot shows a cityscape or not. The aim is to somehow capture the implicit semantics of scene composition.

3.5. Extended 2-stage run (run 1)

Run 1 is based on combinations of different components, namely MPEG-7 Baseline, Concept de-correlation scheme, Face, Body and Motion Descriptors. The MPEG-7 component is based on the following global MPEG-7 descriptors: color layout, edge histogram, homogeneous texture and scalable color (for more details see run 4). The concept de-correlation scheme is actually an extra SVM that is used per concept, using as input the prediction for the concept queried, but also the predictions for the rest of the concepts. Thus, the correlation within the concepts under consideration is taken into account to improve individual classification results. Face detection is performed using combination of the well known approach based on cascades of Haar wavelets 66 and an approach using neural networks 6. Body detection is performed based on histograms of oriented gradients 6. The Motion Activity descriptor is described on the MPEG-7 standard 6. The prediction vectors that have been used for the calculation of the motion activity are extracted from the compressed domain (MPEG-1 & 2), after performing a median 3x3 block filtering which eliminates outlier vectors. Mean value, variance and median of the motion activity within a shot are considered as the descriptor of the shot.

Combining these components, four sub-runs have been created: MPEG7 Baseline, MPEG7 Baseline plus concept de-correlation scheme, MPEG7 Baseline + concept de-correlation scheme + late fusion with Faces & Bodies detections, and Baseline + concept de-correlation scheme + late fusion with Faces & Bodies and shot Motion Activity. The four above sub-runs are ranked for each concept using as quality measure the average precision. The average precision estimation has

been performed considering the predicted folds after making a 10-fold over the annotated content provided by TRECVID. The number of retrieved documents, as is proposed on the contest, is limited to 2000.

It is important to point out that, although theoretically, the separation margin obtained by the SVM should be more precise the more information the input patterns contain, there are various degrees of liberty with could change this behavior. On our case, we have to estimate the best RBF kernel parameters of the SVM using a limited number of combinations of them. Moreover, due to the unbalanced sample sets (*i.e.* much more negative than positive items), a selection of negative items or a replication of positives need to be performed, as is pointed out on 6.The problems come up when the separation capacity of the input patterns is not very high, as on our situation. These reasons make necessary the use of a ranking method on the training process that permits the estimation how good the test results will be, and consequently choose the most proper solution for each concept not only considering how descriptive the input patterns are.

3.6. Combined SIFT & MPEG-7 (run 2)

Run 2 is based on a combination of MPEG-7 and SIFT-based 6 global image features. A set of MPEG-7 features are concatenated to form a single feature vector for each image; the considered MPEG-7 features include color structure, color layout, edge histogram, homogeneous texture and scalable color. In parallel to this, a SIFT-based feature vector is also created for each image in a 2-stage procedure. A set of 500 keypoints, on average, is extracted from each image and a SIFT descriptor vector (with 128 elements) is computed for each keypoint 6. A set of 100 Visual Words is subsequently created by performing clustering in the 128-dimensional feature space and, using the “Bag of Words” (BoW) methodology 6 and the created Visual Words, a new 100-dimensional feature vector is created for the image based on its original SIFT descriptor vectors.

At the first stage of the high-level feature extraction process, the MPEG-7 and BoW feature vectors are exploited independently from each other. A set of SVM classifiers (implemented using LIBSVM 6), comprising one classifier per high-level concept, is trained using the MPEG-7 feature vectors extracted from the first half of the development set. A second set of SVM classifiers is trained similarly using the BoW feature vectors instead of the MPEG-7 ones. In both cases, a subset of the negative samples included in this first half of the training set is selected by a random process and is employed instead of their complete set, in a 5:1 proportion to the positive samples available for each high-level feature, to facilitate the training of the classifiers in cases where the number of positive samples available for training is disproportionately low. The SVM parameters were set by an unsupervised optimization procedure that is part of the LIBSVM tool. The output of classification for an image, regardless of the employed input feature vector, is a number in the continuous range [0, 1] expressing the Degree of Confidence (DoC) that the image relates to the corresponding high-level feature.

Using these classifiers trained on the first half of the development set, the second half of the development set is processed and corresponding DoCs are extracted for its members. The set of formed DoC vectors (one per image and high-level concept) serves as a training set for another set of SVM classifiers, comprising again one classifier per high-level concept, that realize a second stage of classification. The training of them (i.e. selection of a subset of negative samples; optimization of parameters, calculation of output DoC, etc.) is performed similarly to the previous cases. Having completed all the training processes, MPEG-7 and BoW feature vector extraction followed by the first and second SVM classification stages are performed on the test dataset. The results

per high-level feature (DoCs of the second stage SVM classification) are sorted by DoC in descending order, and the first 2000 shots are submitted to NIST.

4. Results

With the exception of run 2 (the combined SIFT/MPEG-7 run), which performed similar to the median results for all participants, the results produced by the submitted runs were much lower than the median, obtaining very low in terms of the mean inferred Average Precision (Mean inferAP), both globally and per feature. The following table summarizes the results:

Run #	**4**	**5**	**3**	**6**	**1**	**2**
Run name	MPEG-7 baseline	SIFT BoW	Visual words	2-stage	Extended 2-stage	2-stage SIFT MPEG7
Total true shots	649	414	724	277	252	1431
Mean inferAP	0.011	0.004	0.013	0.003*	0.002*	0.043

* After the correction of a bug related with the ratio of selected positive and negative samples on the prediction phase, the recomputed mean inferred average precision 6 , on runs 6 and 1 is, respectively, 0.0115 and 0.0131.

MPEG-7 baseline (run 4)

The baseline run performs significantly below the median results for all concepts in terms of Mean inferAP except for feature 9 (*Driver*), and for the set 1-2-3-5-8-20 (though in this last set it is not too difficult to be as good as the median, since the median results are very close to zero).

An important contribution to this low performance was an erroneous implementation that prevented training really optimized for MAP value, due to an error in the computation of MAP[1] Firstly we have discovered that during the training phase the precision values were computed after each relevant shot only if the shot was also predicted to be relevant by the trained classifier (that is, false negatives were not counted). Furthermore, relevant shots that were not retrieved were not included in MAP computation. The above bugs resulted in classifiers trained to rank very highly only a subset of all relevant shots. In fact, this appears to be consistent with the final evaluation results, where the runs affected by the above problems obtained relatively high precision values at the top of ranked lists.

Values for pure precision are thus more encouraging: mean precision at 5 shots is 20%, and 15% for the precision at 10 shots. However the performance varies a lot among features. There are essentially two groups (with some intra-group variations):

- In some of them the values for P5, P10 are low, but significant. These cases get spoiled when we increase the number of results considered; it means that the system has some idea of what it is searching for, but it gets only the “easy cases” (or, considering the above bugs, the cases it was trained to detect). This happens for features 4, 6, 7, 9, 10, 15, 16, 17, 18, 19
- In the other group P5, P10 is zero, and it is only around P100 or so when precision starts to be nonzero (albeit obviously very low). In this case the system is mostly clueless, and only by getting more and more results it

[1] Unfortunately, due to the use of the Baseline in subsequent runs, this error propagated to them.

ends up by including a few right ones. The "clueless" cases are 1, 2, 3, 5, 8, 11, 12, 13, 14, 20

SIFT Bag of Words (run 5)

We believe that the performance of this run suffered from an inadequate choice of several parameters. In the future we plan to investigate further the influence of different methods for selecting key-points, number of keypoint selected in every image, size of the visual word vocabulary and inclusion of spatial information.

Visual words (run 3)

This run contains the baseline with 6 concepts changed (using the Visual Words approach): 1, 6, 7, 10, 16 and 17. Of those 6, the only ones in which run 3 appears to provide an improvement in AP over the baseline are concept 7 (*Two people*) and 17 (*Nighttime*). In concepts 10 & 16 the AP decreases (slightly), in concepts 1 & 6 the AP is about the same.

2-stage classification run (run 6)

As already mentioned, the obtained results for this run were poisoned by a bug when selecting the number of negatives samples to be used on the training of the models that later were going to be used to perform the final predictions. After correcting the bug and repeating the evaluation methodology 6, the obtained mean inferred Average Precision was 0.0115, slightly over the Baseline run results, whose predictions were the input of the second stage in this run (consequently, the target of this run was to improve it).

Extended 2-stage run (run 1)

This run consists on an extension of the previous 2-stage classification run, and presented the same bug as it. Analogy, the mean inferred Average Precision was recalculated 6, finally obtaining an inferAP of 0.0131. Therefore, and as expected, this run now slightly improves the previous run

Combined SIFT & MPEG-7 run (run 2)

This run performs very close to the median for most results. Concepts 6,7,8,9,10,13,15,16,19,20 are slightly above the median, with the rest of the concepts being slightly worse than the median. The combined SIFT & MPEG-7 run is the result of late fusion of two other independent runs, namely a SIFT-based BoW run and an MPEG-7 based one, which were not submitted to NIST due to the limitation of runs to 6. Additional results of experimentation with these two approaches and comparison with the combined SIFT & MPEG-7 approach (also denoted as “fusion” run), in all cases using just the first half of the annotated development set for training and the second half for evaluation (in contrast to the actual run submitted to NIST, where the entire development set was used for training), are shown in Figure 1.

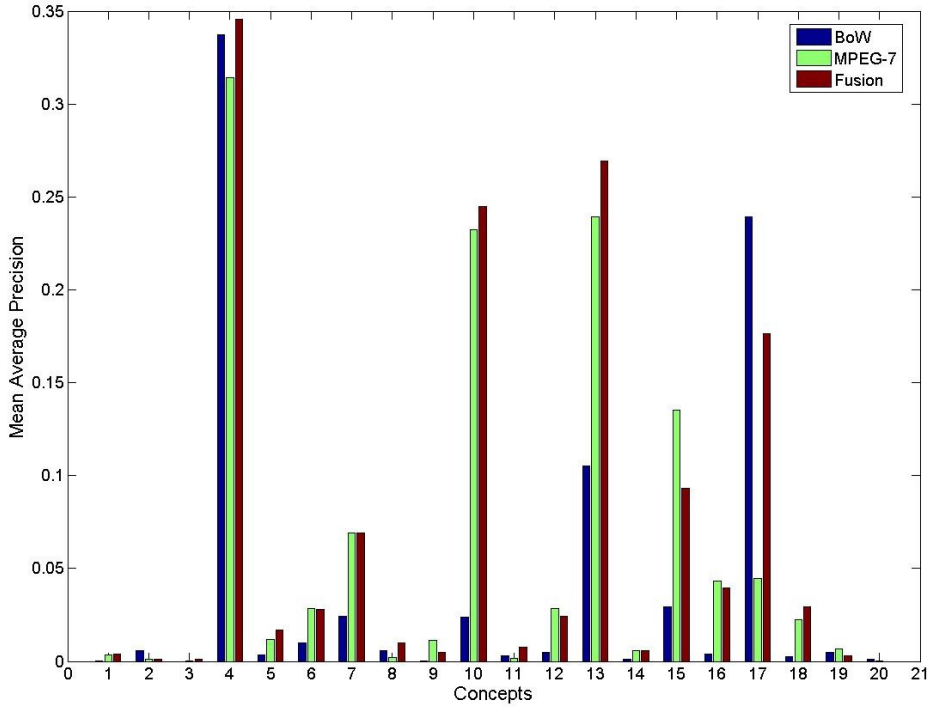

Figure 1 : Mean Average Precision at 100 for BoW, MPEG-7 and combined run.

The combined run is shown to perform better in most cases and even when this is not the case and significant deviations in performance exist between the BoW run and the MPEG-7 one, the combined run manages to perform close to the best performing of the other two in almost all cases.

5. Conclusions

Except for the Combined SIFT & MPEG-7 run (run 2), which performed reasonably, all other runs were poisoned with the implementation bugs included in the MPEG-7 Baseline run[2], and therefore their results are not to be considered definitive of their potential performance; we plan to repeat the training phases with a corrected system.

On relation with 2-stage & Extended 2-stage runs, we plan to do some more work in order to improve the quality of the estimation of performance at development time. This would permit a more proper optimization of configuration parameters of the classification machine (in the case of a SVM, the kernel parameters) and, also, a more intelligent selection of training sets when unbalanced samples are under use.

6. References

[1] A.F. Smeaton, P. Over, W. Kraaij, “Evaluation campaigns and TRECVid”. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722

[2] Run 5 (SIFT BoW) did not use the MPEG-7 Baseline run, but included the same erroneous code for training against MAP values.

[2] C. Petersohn. "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System", TREC Video Retrieval Evaluation Online Proceedings, TRECVID, 2004

[3] P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’01). Hawaii, Dec 11-13, 2001.

[4] H.A. Rowley S. Baluja, T. Kanade, “Neural network-based face detection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 20 (1), page(s): 23-38, 1998

[5] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection”. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). San Diego, CA, USA June 20-25.

[6] B.S. Manjunath (Editor), Philippe Salembier (Editor), and Thomas Sikora (Editor): *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons, April 2002

[7] LK Luo, H Peng, QS Zhang, CD Lin, “A Comparison of Strategies for Unbalance Sample Distribution in Support Vector Machine”, Industrial Electronics and Applications, 2006 1ST IEEE Conference on (May 2006), pp. 1-5

[8] David G. Lowe, "Distinctive image features from scale-invariant keypoints" *International Journal of Computer Vision,* 60, 2 (2004), pp. 91-110. Software available at **http://www.cs.ubc.ca/~lowe/keypoints/**.

[9] J. Sivic, A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos”. In *Proceedings of the Ninth IEEE international Conference on Computer Vision - Volume 2* (October 13 - 16, 2003). ICCV. IEEE Computer Society, Washington, DC, 1470.

[10] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines,2001. Software available at **http://www.csie.ntu.edu.tw/~cjlin/libsvm**

[11] Emine Yilmaz & Javed A. Aslam , “Estimating average precision with incomplete and imperfect judgments”, Proceedings of the 15th ACM international conference on Information and knowledge management, Pages: 102 - 111, Arlington, Virginia, USA, 2006.

[12] P. Kapsalas, K. Rapantzikos, A. Sofou, Y. Avrithis - "Regions Of Interest for Object Detection", Sixth International Workshop on Content-Based Multimedia Indexing (CBMI'08), London, England, 2008.

[13] E. Spyrou, Y. Avrithis - "A Region Thesaurus Approach for High-Level Concept Detection in the Natural Disaster Domain", 2nd international conference on Semantics And digital Media Technologies (SAMT), Italy, Genova, 2007.

[14] M. Everingham, A. Zisserman, C. Williams, L. Van Gool - “The Pascal Visual Object Classes Challenge 2006 (VOC2006) Results”.

[15] E. Spyrou, Ph. Mylonas and Y. Avrithis - "Using Region Semantics And Visual Context For Scene Classification", 1st ICIP Workshop on Multimedia Information Retrieval: New Trends and Challenges October 12, 2008 (Co-located with the ICIP), San Diego, California, USA

[16] G. Quénot, S. Ayache, “Video Corpus Annotation Using Active Learning”, 30th European Conference on Information Retrieval (ECIR'08), pp.187-198, Glasgow, March 30-April 3, 2008

[17] E. Nowak, F. Jurie, B. Triggs, ”Sampling Strategies for Bag-of-Features Image Classification”, ECCV'06.

[18] D. Nister, H. Stewenius, “Scalable Recognition with a Vocabulary Tree”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'06.

[19] J. Yang, Y.-G. Jiang, A. G. Hauptmann, C.-W. Ngo, “Evaluating Bag-of-Visual-Words Representations in Scene Classification”, MIR07.