CEA LIST at Trecvid 2012: Semantic Indexing and Instance Search

Nicolas Ballas[*], Benjamin Labbé[†], Hervé Le Borgne[*†], Aymen Shabou[*]
CEA, LIST, Laboratory of Vision and Content Engineering, Gif-sur-Yvettes, France.
{*firstname.lastname*}*@cea.fr*

Abstract

This paper reports the experiments carried out for the semantic indexing (SIN) and the instance search (INS) tasks at TRECVID 2012. For the SIN task, we evaluated two recently proposed features with a simple one-versus-all linear SVM framework. the first one is a motion histogram based on trajectory vectors. The second one is a bag-of-visterm that take into account the spatial consistency of descriptors. In the INS task, we proposed a descriptors based on local descriptor matching able to scale to the considered corpus. A second contribution for INS consisted in studying several late fusion schemes. Preliminary experiments were conducted on the INS_11 corpus to choose the best strategy, leading to results in the top 5% of past results. While these preliminary results were very promising, 2012 results are above the median of participating runs, but far from reproducing previous year performances. The significance of the results is thus studied, showing that significant difference between two runs is not strictly correlated to the sorted average scores.

1 Semantic Indexing

1.1 Introduction

The TRECVID 2012 semantic indexing (SIN) task [1] is described in the TRECVID 2012 overview paper [2]. We evaluated two specific features recently proposed [3, 4] in the context of the task. In these notes, we report the scores obtained through used a very simple classification scheme (one-versus-all linear SVM). The features have also been proposed in the context of the IRIM participation [5] .

[1] actively participated to the SIN task
[2] actively participated to the INS task

1.2 Feature extraction

Two different low-level feature are investigated in the Semantic Indexing task. First we consider the video motion information with a bag-of-point-trajectory. We also take into account the video appearance through the bag-of-visterm.

1.2.1 Bag-of-point-trajectory

These features captures motion information about local patches motions in videos. Two steps are required to construct the point-trajectory extraction and their aggregation.

Following [6], dense trajectories have been used as local features. Keypoints are densely sampled at multiple spatial scales in each of video frames. Dense optical flow is used to match a point from a frame f to the next frame $f+1$. Trajectories are built by accumulating point correspondences over successive frames. At each frame, a new trajectory is started on a keypoint p if no trajectory is present in a neighborhood.

We characterize the motion using motion histograms based on trajectory vectors, as introduced in [3]. Let $\mathbf{t} = \{(x_1, y_1), ..., (x_L, y_L)\}$ be a trajectory of size L where (x_i, y_i) denotes the point position at frame i. We consider the trajectory motion vectors $\{\mathbf{m}_1, ..., \mathbf{m}_L\}$, where $\mathbf{m}_i = (\mathbf{p}_{i+1} - \mathbf{p}_i)$ and $\mathbf{p}_i = (x_i, y_i)$. These vectors are known to be translation invariant. To achieve scale invariance, they are normalized according to the trajectory maximum vector magnitude. They are also quantized, using their polar coordinates [7], to increase

their robustness toward noise. The quantized motion vectors distribution is captured through an histogram leading to a descriptor capturing the motion information in the trajectory.

Bag-of-words (Bow) model [8] is then used to transform the variable number of trajectory motion descriptors into a fixed-length vector. In this paradigm, a video signature is obtained first by encoding the descriptors according to a learned codebook, and by pooling the obtained codes to end up with a fixed length vector. Saliency coding [9] along max pooling are used to construct the final Bow signature.

1.2.2 Bag-of-visterm on keyframes

The Bag-of-Visual-Words (BoVW) approach [8, 10] is a state-of-the-art representation for visual content description used in image classification. Extended to image description, the usual BoVW design pipeline consists of learning a codebook from a large collection of local features extracted from a training dataset, then creating the global feature of visual signature through coding, pooling and spatial layout. Recent works addressing this problem [11, 12, 13, 14, 15, 4] proved the importance of tuning each of these steps to improve scene classification and object recognition accuracy on different benchmarks.

The pipeline we used is as follows:

- **Local visual descriptors:** dense SIFTs of size 128 are extracted within a regular spatial grid and only one scale. The patch-size is fixed to 16×16 pixels and the step-size for dense sampling to 6 pixels;
- **Codebook:** a visual codebook of size 1024 is created using the K-means clustering method on a randomly selected subset of SIFTs from the training dataset.
- **Coding/pooling:** for coding the local visual descriptors SIFTS, we also fix the patch-size to 16×16 pixels and the step-size for dense sampling to 6 pixels. Then for the extracted visual descriptors associated to one image, we consider a neighborhood in the visual feature space of size 5 for local soft coding and the softness parameter β is set to 10. The max-pooling operation is performed to aggregate the obtained codes and a spatial pyramid decomposition into 3 levels $(1 \times 1, 2 \times 2, 3 \times 1)$ is adopted for the visual-signature. The weight is the same on each pyramid level.

Thus, the size of the visual-signature is equal to $1024 \times (1 + 2 \times 2 + 3 \times 1) = 8192$.

We also tried the process proposed in [4] that modify the coding strategy to take into account the spatial consistency of descriptors. Basically, it forces SIFT descriptors that are close in the image domain to be coded on similar codewords.

1.3 Classification

A one-versus-all linear kernel based Support Vector Machine (SVM) classifier is used, since it has shown good performances in scene categorization task when paired with the max-pooling operation on local features [14, 15].

1.4 Evaluation of submitted runs

Four runs have been submitted, which the results are summarized in table 1

Run name	Description	F results	L results
`CEALIST_1`	BoV_{soft}	0.1137	0.1317
`CEALIST_2`	BoV_{lscr}	0.1024	0.1119
`CEALIST_2`	BoV_{soft} + Motion	0.0820	0.1179
`CEALIST_4`	Motion	0.0070	0.0139

Table 1: Description of the runs submitted and the results ($infAP$) on the full (F) and light (L) task

2 Instance Search

2.1 Introduction

Tee Instance Search Task consist of finding video segments of a certain specific person, object, or place, given a visual example [2].

We used a three descriptors: $Markrs$ is based on local feature matching (section 2.2.1), $Hist_{HSV}$ is a simple HSV color histogram (section 2.2.3) and BoV a bag of visterm similar to the one used in the SIN task (section1.2.2). The retrieval on each descriptor was performed with a naive L_1 distance

based kNN. Retrieval scores were normalized according to different strategies (section 2.3). Finally, we proposed several fusion schemes to do so (section 2.4. Given the results, we evaluated their significance with a Wilcoxon signed-rank test.

2.2 Feature extraction

2.2.1 Markrs

The Markrs are local features for geometrical registration of objects in couple of images. The Markrs process of image description and matching follows the well-known framework of keypoint matching described in [16, 17]. For this experiment, we used the SURF scheme [18] to detect salients keypoints and compute corresponding descriptors, but other descriptors may be used as well within the process described below. They are normalized with respect to their self scale and local orientation of gradient. Then the SURF description is quantized from 64 real values in $(-1, 1)$ into integer values in $[0, 255]$. This leads to a compact description for each keypoint in less than 80 bytes (including 64 bytes for the descriptor).

The image matching process includes two filtering step to drop keyframes of the database that are not close enough to the query. The first filtering step finds matching keypoints with respect to their appearance in a query-candidate couple of images. Valid keypoint matches are considered if they pass the test of relative nearest-neighbors proposed by D.Lowe in [16]. The images with the highest number of matches are top ranked.

The second filtering step selects within the previous results those that provides a similar geometrical configuration of keypoints in the query-candidate couple of images. We avoid considering complete homographies, preferring simple similarities that are much fastest to compute. This reduces the complexity of the exhaustive test of models for this geometrical confirmation. Hence, even a small set of matching keypoints between two images can lead to a fit. The final result list is composed of images having more than p keypoints fitting the geometrical model ($p \geq 5$).

This matching process can detect the co-occurrence of small objects in a query-candidate couple of images, leading to relative good precision for CBIR tasks similar to instance search or duplicate-detection.

2.2.2 Bag-of-visterm

We used the same process as the one described in section 1.2.2.

2.2.3 Global features

We used a well-known color histogram for global image description. The color histogram counts the occurrences of 162 shades in the HSV color space. The similarity between two images is measured as the inverse of a *dLog* distance between two histograms, defined as: [19]:

$$dLog(q, d) = \sum_{i=0}^{i<M} |f(q[i]) - f(d[i]) \qquad (1)$$

$$f(x) = \begin{cases} 0, & \text{if} \quad x = 0 < \alpha \\ 1, & \text{if} \quad 0 < x \leq 1 \\ \lceil \log_2 x \rceil + 1 & \text{otherwise} \end{cases} \qquad (2)$$

Where q and d are two histograms with M bins and $\lceil . \rceil$ is the ceilling function.

2.3 Score normalization

Raw scores for individual feature extractor have various amplitude. Thus, we tested several normalization schemes of these scores, previously used for multilingual information retrieval data fusion [20].The considered normalization schemes are presented in Table 2. The $norm_{man}$ scheme normalizes the score by the maximal score s_{max} obtained for this query. The $norm_{lin}$ scheme linearly normalizes scores between 0 and 1. The $norm_{gauss}$ scheme is equivalent to a Gaussian normalization of scores added to an offset such that they are positive.

$norm_{man}$	s'_i	$=$	$\frac{s_i}{s_{max}}$
$norm_{lin}$	s'_i	$=$	$\frac{s_i - s_{min}}{s_{max} - s_{min}}$
$norm_{gauss}$	s'_i	$=$	$\frac{s_i - s_{min}}{s_\sigma}$

Table 2: Normalization schemes for the scores in the merging strategy

2.4 Fusion Schemes

A topic is represented by n image queries, themselves described according to p descriptors. Hence, we have to merge $n \times p$ lists of results for a given topic. However, considering the $N \times P$ lists of results at the same level is a *global fusion scheme*. In this paper, we studied the effect of two alternative fusion strategies, further named *query-first fusion* and *descriptor-first fusion*. These two-steps fusion schemes consist on merging the result lists according to queries (resp. descriptors) first, then merging the resulting lists into a unique final one (figure 1).

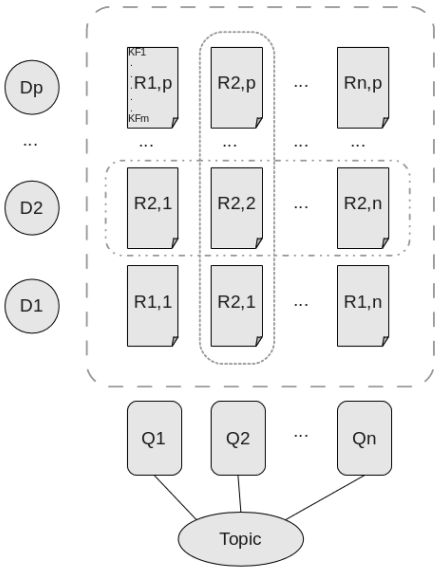

Figure 1: Fusion of results for a given *topic* defined by several *queries* $(Q_1, Q_2, \ldots, Q_n)$, according to several *descriptors* $(D_1, D_2, \ldots, D_p)$

Let consider two lists of results, ordered according to a their scores. They can be merged according to several operator [21, 22], acting directly on the scores or on the rank only. The operators we considered are given in table 3

In our case, the fusion operator may not be the same at each fusion level.

CombSUM	s_f	$=$	$\sum_{i=1}^{L} s_i$
CombMAX	s_f	$=$	$\max_{i=1}^{L} s_i$
CombMEAN	s_f	$=$	$\frac{1}{L} \sum_{i=1}^{L} s_i$
CombMNZ	s_f	$=$	$L * \sum_{i=1}^{L} s_i$

Table 3: Combination operator to merge L lists of results. They can act on the score or the rank.

2.5 Preliminary experiments

Experiments were conducted on the INS 2011 corpus to determine which strategies were the most efficients. Globally, the use of two descriptors can lead to better results when an appropriate fusion strategy is used. However, most of the time, it can result into an average results from those obtained with individual descriptors (thus worse than the best of them). Several results of global fusion strategy, leading to an improvement over the baseline alone (Markrs), are reported in table 4.

Descriptor + weight	Combination operator	Results (mAP)
$H_{HSV}(g)$	CombMAX (s)	23.0
$H_{HSV}(g)$	CombMAX (r)	23.6
BoV (gauss)	CombMAX (s)	26.6
Markrs	CombMNZ (s)	36.9
$1.2 \times Markrs$ + BoV	CombMAX (r)	38.8
$300 \times Markrs$ + BoV	CombMNZ (r)	38.1
$300 \times Markrs$ + BoV	CombSUM (s)	40.6
$40 \times Markrs$ + BoV	CombMAX (s)	40.4
$100 \times Markrs$ + BoV(g)	CombSUM (s)	40.7

Table 4: Results on INS 2011, with a global fusion (one-step) strategy. Combination operator is applied on scores (s) or ranks (r). Descriptors may be normalized with a Gaussian (g) scheme (see table 2)

Results with a *query-first* fusion strategy is reported in table 5. The Markrs list is obtained with the *CombSUM* operator applied on scores.

The Gaussian normalized BoV ($BoV(g)$) list is obtained with the $CombMAX$ operator applied on scores. The Gaussian normalized color histogram ($H_{HSV}(g)$) list is obtained with the $CombMAX$ operator applied on ranks. The Markrs list is weighted to be preponderant, ten times larger than others. The use of the $CombMEAN$ operator leaded to poor results, not reported here. Globally this two-steps fusion strategy result in better results for all operator, whether applier to scores or ranks. Moreover, these results are quite the same for all these operator

Some results with a *descriptor-first* strategy in table 6. All descriptors are first merged to obtained a unique list for each query, using a first combination operator. For this operation, the Markrs descriptor is weighted 100 times mores than other descriptor, such that its results are preponderant. Then, a second fusion is performed on the query lists, using a second combination operator. Results are almost the same as those obtained with the query-first strategy.

Descriptors	Combination operator	mAP
$10 \times Markrs$ + BoV(g)	CombSUM(s)	41.4
$10 \times Markrs$ + BoV(g)	CombSUM(r)	41.1
$10 \times Markrs$ + BoV(g)	CombSUM(s)	41.4
$10 \times Markrs$ + BoV(g)	CombSUM(r)	41.2
$10 \times Markrs$ + BoV(g)	CombMAX(s)	41.2
$10 \times Markrs$ + BoV(g)	CombMAX(r)	41.2
$10 \times Markrs$ + BoV(g)+ $H_{HSV}(g)$	CombSUM(r)	44.5
$10 \times Markrs$ + BoV(g)+ $H_{HSV}(g)$	CombMAX(r)	44.4
$10 \times Markrs$ + BoV(g)+ $H_{HSV}(g)$	CombMNZ(r)	44.3
$100 \times Markrs$ + BoV(g)+ $H_{HSV}(g)$	CombSUM(r)	45.2

Table 5: Results on INS 2011, with a query-first strategy. Combination operator is applied on scores (s) or ranks (r).

Run name	Global description	Results
`CEALIST_1`	Query-first fusion	0.1216
`CEALIST_2`	Markrs alone	0.1135
`CEALIST_3`	Descriptor-first fusion	0.0269
`CEALIST_4`	Query-first fusion	0.1215

Table 7: Description of the runs submitted and the results (mAP) on the INS track.

2.6 Evaluation of submitted runs

Four runs have been submitted, as summarized in table 7. In details, the runs have been constructed as following:

- **CEALIST_2** is our baseline giving the results from the Markrs descriptor only. Queries are merged with $CombSUM$.

- **CEALIST_1** is a query-first fusion of the three descriptor considered. BoV and HSV histogram are normalized according to the $norm_{gauss}$ scheme then merged with the $CombMAX$ operator. The three lists are then merged with $CombSUM$ on the rank (first 500 only), using an equal weight for BoV and HSV histogram and a predominant weight for Markrs.

- **CEALIST_4** is another query-first fusion of the three descriptor considered. BoV and HSV histogram are normalized according to the $norm_{gauss}$ scheme then merged with the $CombMAX$ on the rank (limited to 2000 for BoV and 5000 for $Hist_{HSV}$). The three lists are then merged with $CombMNZ$ on the rank (first 500 only), using an equal weight for BoV and HSV histogram and a predominant weight for Markrs.

- **CEALIST_3** is a descriptor-first fusion of the three descriptor considered. The descriptors are first merged through $CombMAX$ for each query then the resulting lists are merged with $CombMNZ$ to get the final topic list.

A Wilcoxon signed-rank test (null hypothesis is *median difference between the pairs is zero*) shows that the difference is significant between `CEALIST_1` and `CEALIST_2` (p-value 0.000438), as

Descriptors	First combination	Second combination	mAP
Markrs + BoV(g)	CombMAX(s)	CombMNZ(s)	40.6
Markrs + BoV(g)	CombMAX(s)	CombSUM(s)	41.1
Markrs + BoV(g) + $H_{HSV}(g)$	CombMAX(s)	CombMNZ(s)	43.9

Table 6: Results on INS 2011, with a descriptor-first strategy. Combination operators is applied on scores (s) or ranks (r).

Run name	Score	CEALIST_1	CEALIST_4	IRIM_2	IRIM_4	IRIM_1	IRIM_3	CEALIST_2
CEALIST_1	0.1216	x						
CEALIST_4	0.1215	0.88	x					
IRIM_2	0.1192	0.71	0.66	x				
IRIM_4	0.1173	0.06	0.10	0.83	x			
IRIM_1	0.1171	**<0.01**	0.08	**<0.01**	0.15	x		
IRIM_3	0.1162	**<0.01**	0.07	0.68	0.06	0.08	x	
CEALIST_2	0.1135	**<0.01**	**0.02**	0.46	**<0.01**	**<0.01**	**<0.01**	x

Table 8: p-value resulting from a Wilcoxon signed-rank test for all our runs and those of IRIM [5]. Significant p-values (below 0.05) are shown in bold. Those displayed as **<0.01** are very small and thus significant.

well as between CEALIST_4 and CEALIST_2 (p-value 0.019809), but not between CEALIST_1 and CEALIST_4 (p-value 0.878851).

Contrary to the preliminary experiments, the *Descriptor-first fusion* did not lead to good results. It may be due to the particular nature of the 2011 corpus, composed of quite coherent videos.

Last, we compared significance of results for our runs and those of IRIM [5] and reported the p-value in table 8 (significant ones are bold). It shows that the two best results of both teams are not significantly different, while their scores range from 0.1173 to 0.1216. More surprising the two last runs of IRIM (IRIM_1 and IRIM_3) are significantly less than CEALIST_1 but not CEALIST_4, while these two runs are almost the same (similar score and not significantly different between them). However, the most surprising result is that the best IRIM run (IRIM_2) is the only one that is not significantly different from the baseline. To explain this result, we computed the mAP difference between the baseline and each IRIM run (*i.e* the quantities computed in the Wilcoxon test) and plotted these values in ascending order (Fig. 2). Hence, we can see that IRIM_2 is both strongly better and strongly worse than the baseline, while other IRIM runs are always better or similar to the baseline.

Most important points we retain from the INS 2012 experiments are thus:

- The two steps fusion scheme lead to better results than the global fusion strategy
- However, the descriptor-first strategy can fail with some heterogeneous corpus. The query-first strategy is efficient in this case.
- All fusion operator give good results in a query-first strategy, both with scores and ranks, except $CombMEAN$.
- The actual significance of the results should be studied carefully. The significant differences between two runs is not strictly correlated to the sorted average scores. In other words, when runs are sorted according to their average score (*e.g* mAP), significant difference can be intercalated between non significantly different runs.

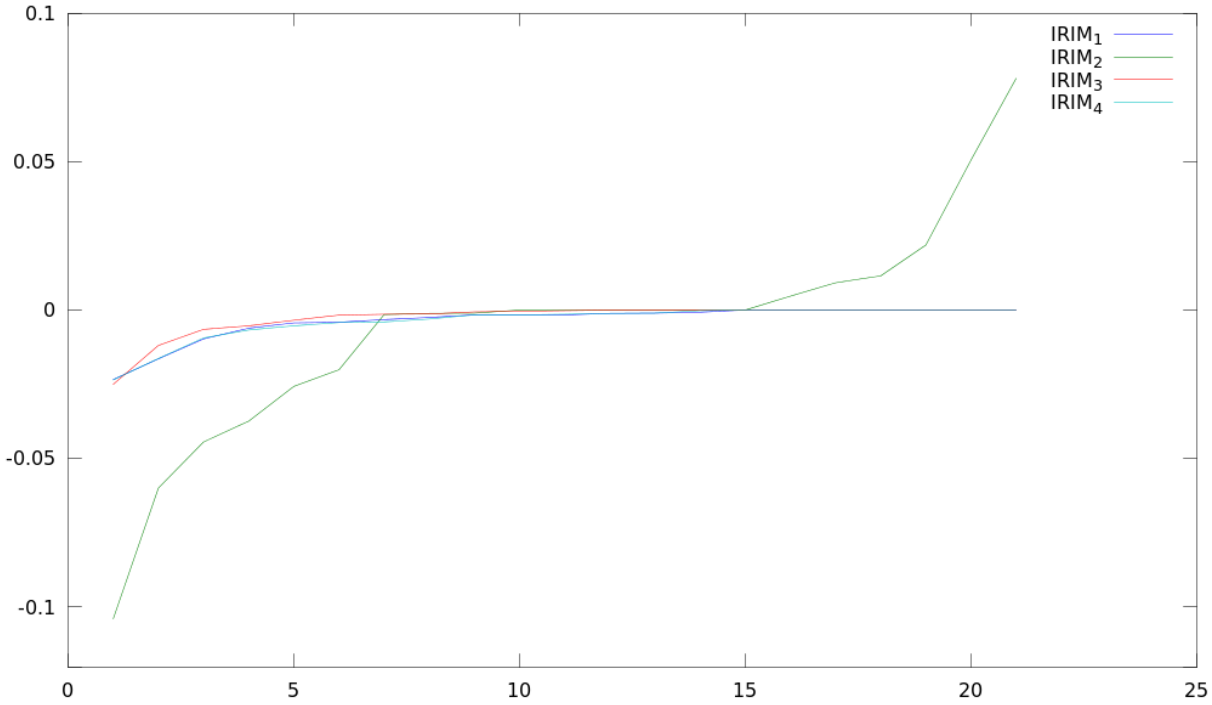

Figure 2: Sorted mAP differences of IRIM runs with the baseline (CEALIST_2). While most of IRIM runs are always better or same as the baseline, the run IRIM_2 is both strongly better and worst than CEALIST_2.

3 Acknowledgements

Experiments presented in this paper (for the SIN task) were carried out using the Grid5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

We were part of the IRIM consortium (a GDR-ISIS research network from CNRS; see [5] for a presentation of the extensive work) that provided a rigorous organisation of Trecvid data on Grid 5000 as well as handy scripts and programs to carry out the experiments and, above all, organized a large scale evaluation of the methods on past Trecvid data.

First fusion experiments in INS task were carried out with the program of Boris Mansencal from LABRI.

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications* (A. Divakaran, ed.), pp. 151–174, Berlin: Springer Verlag, 2009.

[2] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot, "Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*, NIST, USA, 2012.

[3] N. Ballas, B. Delezoide, and F. Prêteux, "Trajectories based descriptor for dynamic events annotation," in *ACM workshop on Modeling and representing events*, pp. 13–18, ACM, 2011.

[4] A. Shabou and H. Le Borgne, "Locality-constrained and spatially regularized coding for scene categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625, 2012.

[5] N. Ballas, B. Labbé, A. Shabou, H. Le Borgne, P. Gosselin, M. Redi, B. Merialdo, H. Jégou, J. Delhumeau, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, F. Thollard, N. Derbas, G. Quénot, H. Bredin, M. Cord, B. Gao, C. Zhu, Y. tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benot, P. Lambert, T. Strat, J. Razik, S. Paris, H. Glotin, T. Ngoc Trung, D. Petrovska Delacrétaz, G. Chollet, A. Stoian, and M. Crucianu, "IRIM at TRECVID 2012: Semantic Indexing and Instance Search," in *Proc. TRECVID Workshop*, (Gaithersburg, MD, USA), nov 2012.

[6] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169–3176, IEEE, 2011.

[7] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009.

[8] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 1470–1477, Oct. 2003.

[9] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 1753–1760, IEEE, 2011.

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

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.

[12] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, 2009.

[13] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2559–2566, 2010.

[14] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, 2010.

[15] L. Liu, L. Wang, and X. Liu, "In Defense of Soft-assignment Coding," in *IEEE International Conference on Computer Vision*, 2011.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, vol. 2. Cambridge Univ Press, 2000.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[19] R. Stehling, M. Nascimento, and A. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 102–109, ACM, 2002.

[20] R. Besançon and C. Millet, "Data fusion of retrieval results from different media: Experiments at imageclef 2005," in *Accessing Multilingual Information Repositories* (C. Peters, F. Gey, J. Gonzalo, H. Mller, G. Jones, M. Kluck, B. Magnini, and M. de Rijke, eds.), vol. 4022 of *Lecture Notes in Computer Science*, pp. 622–631, Springer Berlin / Heidelberg, 2006.

[21] E. Fox and J. Shaw, "Combination of multiple searches," *NIST special publication*, pp. 243–243, 1994.

[22] N. Belkin, P. Kantor, E. Fox, and J. Shaw, "Combining the evidence of multiple query representations for information retrieval," *Information Processing & Management*, vol. 31, no. 3, pp. 431–448, 1995.