Retina-Enhanced SURF Descriptors for Semantic Concept Detection in Videos

Tiberius Strat, Alexandre Benoit, Patrick Lambert, Alice Caplier

Retina-Enhanced SURF Descriptors for Semantic Concept Detection in Videos

Sabin Tiberius STRAT[1,2], Alexandre BENOIT[1], Patrick LAMBERT[1] and Alice CAPLIER[3]
[1] LISTIC - Université de Savoie
Annecy Le Vieux, France
e-mail: Sabin-Tiberius.Strat@univ-savoie.fr, alexandre.benoit@univ-savoie.fr,
patrick.lambert@univ-savoie.fr
[2] LAPI - University "Politehnica" of Bucharest
Bucharest, Romania
[3] Gipsa-Lab - Université de Grenoble
St Martin d'Hères, France
e-mail: alice.caplier@gipsa-lab.grenoble-inp.fr

***Abstract*— This paper proposes to investigate the potential benefit of the use of low-level human vision behaviors in the context of high-level semantic concept detection. A large part of the current approaches relies on the Bag-of-Words (BoW) model, which has proven itself to be a good choice especially for object recognition in images. Its extension from static images to video sequences exhibits some new problems to cope with, mainly the way to use the added temporal dimension for detecting the target concepts (swimming, drinking...). In this study, we propose to apply a human retina model to preprocess video sequences, before constructing a State-Of-The-Art BoW analysis. This preprocessing, designed in a way that enhances the appearance especially of static image elements, increases the performance by introducing robustness to traditional image and video problems, such as luminance variation, shadows, compression artifacts and noise. These approaches are evaluated on the TrecVid 2010 Semantic Indexing task datasets, containing 130 high-level semantic concepts. We consider the well-known SURF descriptor as the entry point of the BoW system, but this work could be extended to any other local gradient based descriptor.**

***Keywords*— Bag of words, Retina analysis, Retina preprocessing, Semantics, SURF, Video content, Video indexation**

I. Introduction

The ever increasing abundance of digital multimedia content demands applications able to organize this content automatically. Such applications, for searching and browsing through large databases, rely on a content-based, semantic annotation. Because of the large volume of data, this annotation can only be done automatically.

Following this idea, this paper explores the task of detecting semantic concepts automatically in short fragments of video sequences (called *video shots*). Many approaches have already been explored on single images [7], while some are also delving into the more challenging task of analyzing videos [13]. Within this topic, one of the most popular approaches is the so-called Bag of Words (BoW) model [5].

The basic idea of the BoW model, illustrated in Fig.1, consists in first extracting local features from the images or video frames using appropriate descriptors (such as the well-known SIFT or SURF descriptors). The next step is to cluster these features into a dictionary of "visual words", using a clusterisation algorithm. Afterwards, the images or videos are represented as histograms of visual words (called Bags of Words, BoW), which encode the rate of appearance of these visual words in the image or video. In the end, a supervised classification algorithm is employed in order to distinguish between the BoW representing the target concept (also called High Level Feature, HLF) and the other cases.

Depending on the application, choices need to be made regarding which features to use (SIFT, SURF etc.), how to sample the features from the scene (dense grid, interest point detector etc.), how to cluster the features into dictionaries (hierarchical clustering, centroid-based, density-based etc.) and which classification algorithm to use (Support Vector Machines with various kernels, K-Nearest Neighbors etc.). With so many variables, the complete toolchain is difficult to handle, but good recipes have been proposed and compared [7], [14], and the performances on static images, especially for object detection, are generally good.

The next difficult step is using the added temporal dimension inherent to videos. The TrecVid challenge [12] illustrates the difficulty of the problem. It aims at encouraging innovation in the specific domain of video databases exploration with various contest tasks, of which our team was interested in the High-Level-Feature detection (the Semantic indexing task). In this competition, videos come from extremely varied sources, from quality broadcast TV to very low quality smartphone videos with ever-changing resolution, compression ratio, camera setup, handling, lighting conditions etc. With such non-homogeneous data, robust yet meaningful descriptors are needed. Many participants in TrecVid employ classical SIFT, SURF and other low-level image descriptors for constructing BoW histograms, generally focused on specific keyframes extracted from the video subsequence. Indeed, working only on keyframes helps keep the computational costs down and also allows the use of already proven methods for static images. However, temporal information is ignored. Some

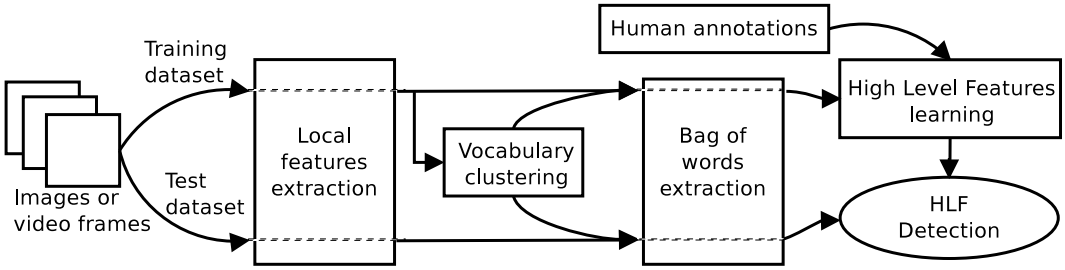

Fig. 1: State-of-the-art performing Bag of words processing toolchain for High Level Features detection.

solutions to consider the temporal aspect have already been proposed, such as MoSIFT [4], which extracts features solely on blobs containing motion. Some methods choose to describe motion by computing histograms of optical flow [6], while others first detect spatio-temporal interest points and then characterize these points with various spatial (histogram of oriented gradients) and motion (histogram of optical flow) descriptors [11]. There are also methods which are based on describing the trajectories of interest points in videos, such as [15] and [2]. Nevertheless, with all these methods there is a risk that static scenes will not be described properly, because of the lack of movement.

In this paper, we focus on the problem of developing a descriptor which should be robust to artifacts typical of video streams (noise and compression effects) and to typical lighting problems (quantity and quality of light, white balance and, if possible, light direction and shadows). Our aim is to clean the video content from disturbances and enhance the relevant information in order to increase the extracted descriptor's discriminative power. We choose an efficient and well known spatial descriptor: the SURF descriptor, and, keeping the same BoW-based processing toolchain, we add a video preprocessing step which uses the retina model from [3]. This model presents interesting properties for filtering out undesired image artifacts and gaining robustness to luminance variations. More precisely, the parvocellular output of the retinal model allows enhancement of details and artifact suppression. We conduct a performance evaluation of such preprocessing and compare results using the inferred average precision metric used in the TrecVid challenge [17], [16].

The remaining of the paper is structured as follows: section II describes the retinal preprocessing, section III presents the evaluation protocol and section IV compares the results of the classical approach and of our approach, before drawing conclusions.

II. Retinal Preprocessing

When facing the problem of heterogeneous video datasets, the low-level features being extracted should preserve their consistency across the various contents. Thus, compression, light changes, noise effects and background clutter should be minimized before feature extraction (Bag Of Words or any other feature). To this end, we propose to use the real-time retina model proposed in [3], briefly described in the following paragraphs.

The human retina generates data channels for motion and spatial details analysis. We focus on the so called *parvocellular* channel which processes spatial details and colors. It has a high resolution in the center of the visual field, where it constitutes the foveal vision. It normalizes colors, enhances local contrast, responds well to temporally-sustained signals, while smoothing out fast temporal variations.

The parvocellular channel improves the image in several ways. It enhances contours of medium spatial frequencies, representing spatial details, while leaving out high-frequency components, related to noise and compression artifacts. It also smooths movement, thereby eliminating high temporal frequencies which are easily corrupted by noise and video compression. An additional benefit is the local adaptation of luminance, which enhances details even in very dark areas of the scene, but without amplifying the image noise. More information about the human retina from a signal processing point of view can be found in [9].

In our method, we implement the parvocellular channel as a sequence of color images with all of the enhancements noted above. An example of the effect of the parvocellular channel on a video can be seen by comparing Fig. 2a with Fig. 2b, where the enhancement can be seen especially around the facial features (mouth, eyes, eyebrows etc.).

In our study, we assume that the resolution across the image is the same even if this is not the case for a real, biological retina. This allows us to get enhanced spatial details information at any position in the considered video sequence. The parameters used in the retina setup correspond to the default values described in [3] and the ones available in the retina code distributed within the OpenCV[1] library.

As illustrated in Fig. 3, we involve the retinal model by preprocessing video keyframes with the retina's parvocellular channel, as described below. We also add, for comparison, the basic keyframe-based SURF descriptor, without any kind of retinal processing.

[1] `http://www.opencv.willowgarage.com`

(a) Input frame (b) Parvocellular channel output

Fig. 2: Retinal parvocellular processing example, after the retina has reached its stable state.

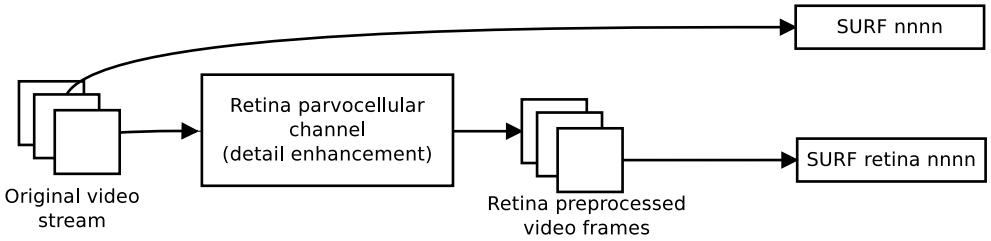

Fig. 3: Proposed preprocessing method: the input video is preprocessed by the retinal parvocellular channel, before extracting classical SURF features on a dense grid. Two descriptors are generated for comparison, one from the original keyframe (*SURF nnnn*, where *nnnn* is the vocabulary size) and one from the parvocellular output at the keyframes (*SURF retina nnnn*). More about these descriptors in section III.

Keyframe Preprocessing

As previously stated, a classical approach for detecting concepts in videos is to analyze only a few *keyframes* extracted from the considered video shots, thereby reducing computational cost greatly. But in our approach, instead of simply analyzing the original keyframes (which would give the descriptor *SURF nnnn* in figure 3, where *nnnn* is the vocabulary size), we analyze the output of the parvocellular pathway at the same moments as the keyframes (giving rise to *SURF retina nnnn*). The reasoning behind this is that since the person shooting the movie generally points the camera towards the region of interest, relevant information tends to remain static on the foveal retina (or the TV screen). This way, the retinal parvocellular channel is able to clean the region of interest of noise or compression artifacts, enforce spatial details and normalize colors, therefore providing the basis for cleaner image descriptors in later stages. As the next stage, we use the classical BoW approach from figure 1, extracting visual features on a dense grid at each video keyframe of the retinal parvocellular output.

From an implementation point of view, in order to avoid the transient response which appears when initializing the retina and which allows only low spatio-temporal frequencies to pass, we actually start the retinal processing 20 frames before the keyframe (after this interval, the response reaches its stable state), but we only collect descriptors at the moment of the keyframe.

Note that regarding the use of temporal information inherent in videos, we do not describe motion or other complex spatio-temporal property of the video stream, we just describe isolated keyframes from the video shots. However, we describe isolated preprocessed keyframes benefitting from the temporal properties of the spatio-temporal retina model, which eliminates high spatial frequencies inherent to noise and compression block effects.

The dense grid setup consists of SURF features extraction (OpenCV implementation of opponent SURF) using a 9 pixels sampling rate on the video frames. Afterwards, we construct the feature vocabulary using the OpenCV implementation of Kmeans clustering. This is performed in 3 passes on the training set, using the Kmeans++ initialisation method [1]. A fixed-length descriptor is generated for each keyframe, comprised of the histogram of visual words, the size of which is determined by the chosen number of clusters (either 1024 or 4096). Afterwards, the Bag of Words classification stage is performed using a K-Nearest Neighbors classifier, as described in [8].

III. Evaluation Protocol

We test our methods on the TrecVid 2010 Semantic Indexing Challenge development dataset, which contains 119685 video shots of short length (between a few seconds up to tens of seconds), on which the presence or absence of 130 various semantic concepts (such as “asian people”, “vegetation”, “cityscape”, “harbor”, “ambulance”, airplane flying”, “throwing”, “cheering” etc.) has been annotated. We split the database in half: 59885 shots are used for extracting the BoW vocabularies and training the classifiers, while the other half is

used for evaluating the performances. TrecVid also provides an official selection of keyframes to represent each of the video shots, and we have used it in our experiments.

We do not experiment with different ways of splitting the database in order to obtain results in cross-validation, for three reasons. First, as the database is very large, this reduces the need for cross-validation, the results being already representative. Second, cross-validation on large datasets requires great computational resources. And finally, in such a way, our experiments comply with the official TrecVid evaluation procedure.

The goal of the Semantic Indexing task is to return, for each of the 130 concepts, a ranked list of shots containing the concept. We use the official TrecVid measure of performance, the *mean inferred average precision* [16], [17], to quantify our results.

In this context, in order to evaluate the improvements obtained with our methods compared to traditional ones, we have constructed, for each video shot, four descriptors:

- *SURF_1024* and *SURF_4096*: SURF features extracted on a dense grid, from the original video keyframe; K-means vocabulary clustering produces 1024 and 4096 visual words (therefore a 1024 respectively 4096 dimensional descriptor);
- *SURF_retina_1024* and *SURF_retina_4096*: similar to previous, just that now we extract the features on the parvocellular output frame instead of the original keyframe;

All the methods use the same BoW extraction toolchain shown in Fig. 1, with the same keyframes, the same sampling rate on the dense grid, the same parameters for the retina and the same parameters for the K-means vocabulary generation. The final concept detection is performed using the KNN approach described in [8], always using the same setup.

As a technical detail, because each concept in the TrecVid database is only present in a very small fraction of the total number of shots, for the vocabulary construction, we have selected a subset of 1008 video shots from the training database, such that at least 25% of the selected training shots contain at least one positive example of any one of the target concepts. This allows vocabulary extraction on 1196705 SURF features. The same subset of shots was used to generate the vocabularies for each of the methods.

IV. Results Analysis

Figure 4 presents the mean inferred average precision on the 130 High Level Features of our test database, obtained with the 4 proposed descriptors. First of all, these results might appear low, but they are in the same range as other BoW approaches for this dataset, and well above the noise level (which is less than 0.001). We must not forget that in the TrecVid datasets, there are many concepts which have very few positive examples, in the order of a few tens of positives, sometimes even less, for several tens of thousands of negatives, which causes the average precisions to be very low.

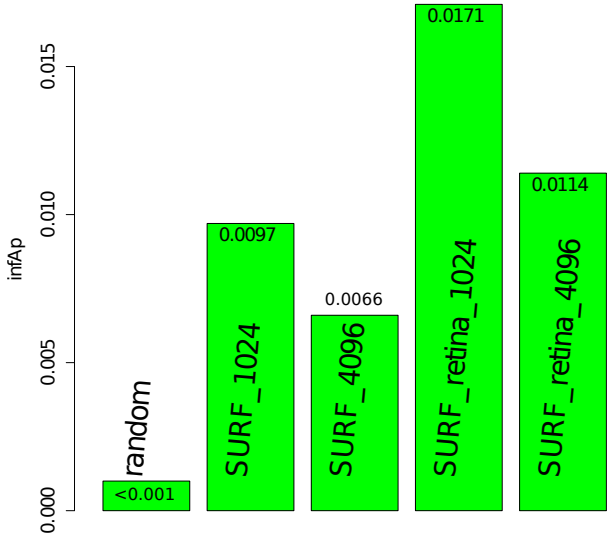

Fig. 4: Mean infered average precisions (infAP) [17] obtained by the compared methods (plus a random classification), on the test set.

For example, [8] obtain mean inferred average precisions in the range of 0.048-0.054 for descriptors based on dense SIFT, on the same database and with the same vocabulary size.

However, our results are not directly comparable with [8], for the following reasons: it is known that SIFT tends to generally give better results than SURF [10], but we chose SURF because of the lower computation time compared to SIFT; additionally, we did not yet optimize the parameters of each step of our approach (the video shots on which to extract the vocabulary, the dense grid setup, the parameters of the retina, the classification algorithm etc.), which is also a reason for our lower performances. To circumvent the problem of result comparison, we generated our own baseline *SURF 1024* and *SURF 4096* descriptors, which used exactly the same parameters as the retina-based methods, for all steps except the retinal processing of course. By fine-tuning the parameters and by fusing more descriptors of various types (various fusion strategies exist, ranging from simple arithmetical fusions, to more sophisticated methods employing hierarchical fusions, feature selections etc.), the performances of the global system could be increased, as can be seen for example in the team submissions of [8] for TrecVid. But, in this paper, our aim is first to show the improvement provided by a fast bio-inspired preprocessing.

A. Global performances

The striking point is that all the methods using the retina outperform the classical SURF descriptors. The retina processed keyframe based descriptor (*SURF_retina_nnnn*) increases performance by 76% (for 1024 visual words), respectively 73% (for 4096 visual words) compared to the classical keyframe-based SURF descriptor (*SURF_nnnn*). Therefore, the parvocellular preprocessing used in *SURF_retina_nnnn*

improves results significantly compared to the baseline.

This performance increase can be explained through the image enhancement brought by the parvocellular channel, thereby allowing a more accurate extraction of details, improving performances for concepts related to specific objects and texture recognition. We also see that a vocabulary size of 1024 is better than 4096, because 4096 fragments the feature space into "too small" visual words, and we become too sensitive to small variations of image appearance.

From a general point of view, this experiment proves that enhancing relevant details, while filtering out image artifacts ranging from noise and compression effects to under/overexposure problems, allows a classical descriptor such as SURF to be significantly improved.

B. Performances concept by concept

We continue our analysis at a concept-by-concept level, in order to see which semantic concepts benefit the most from the retinal preprocessing, and which concepts are penalized by this preprocessing. From a general point of view, we counted 33 out of the 130 concepts for which *SURF retina 1024* was noticeably better than *SURF 1024*. Regarding the remaining concepts, performance differences are not that significant while average precisions remain low. This is explained by the fact that the SURF descriptor with and without preprocessing is less adapted. This follows the idea that a single descriptor cannot be efficient for all the concepts, and this is why fusion strategies between various kinds of descriptors are used in the TrecVid competition.

We present in Table 1 the most representative concepts for comparing the performances of both approaches, 13 of those 33 for which *SURF retina 1024* was better than *SURF 1024*, and all 8 concepts for which the the simpler method was better. A first thing to notice is that whenever *SURF retina 1024* is better than *SURF 1024*, the difference between the two methods is a lot greater than in cases when the simpler method is better, thereby supporting the idea that our preprocessing greatly improves results most of the time, and when it doesn't improve, at least it doesn't degrade performances significantly (this, of course, is translated in the global average precision improvement seen in Fig. 4).

When the retina approach reacts better to concepts, it can be seen that such concepts are related to details and situations where light changes must not be taken into account, but can disturb the classical descriptor a lot. For example, "Beach", "Computer or TV screen", "Crowd", all of these concepts can be acquired in various lighting conditions so that the retina light cleaning effect improves the detection. As a general rule, the added value of the details enhancement effect helps improve the performance for concepts related to spatial structures and textures. On the other hand, some concepts do not benefit from the retinal preprocessing, but this can be explained by the model properties. For example, "Actor" and "Highway" are much better detected without the retina. This can easily be explained by the motion related to these concepts (motion of actors, cars on the highway). Indeed, the retina cancels the mid-spatial frequencies of fast moving objects, but in the case of such concepts, it eliminates an important part of the useful information.

Table 1: Infered average precisions obtained by *SURF 1024* and *SURF retina 1024* on the test set, for some particular concepts.

concept	SURF 1024	SURF retina 1024
Anchorperson	0.0834	**0.2328**
Beach	0.0127	**0.1028**
Cheering	0.0140	**0.0555**
Computer or TV screens	0.0795	**0.1536**
Crowd	0.0008	**0.0189**
Female person	0.0029	**0.0170**
Instrumental musician	0.0081	**0.0283**
Maps	0.0163	**0.0475**
News studio	0.0706	**0.1590**
Nighttime	0.0023	**0.0271**
Reporters	0.0759	**0.1892**
Road	0.0137	**0.0574**
Actor	**0.0134**	0.0066
Bridges	**0.0166**	0.0088
Buildings	**0.0237**	0.0158
Daytime outdoor	**0.0447**	0.0341
Highway	**0.0133**	0.0000
Landscape	**0.0371**	0.0108
Sky	**0.0768**	0.0195
Vegetation	**0.0588**	0.0488

A noticeable performance difference between *SURF 1024* and *SURF retina 1024* is the good performance of the latter approach for the "Nightime" concept, but its lower efficacy for the "Daytime outdoor" concept. This can be explained by the fact that the proposed retina parameters were adjusted in the aim to cancel the mean luminance of the input images. Then, when "Daytime outdoor" has to be detected, the retina does not provide any significant information since the high mean luminance criterion has been cancelled. Even if details related to the outdor concept are reinforced, the brightness information is lost and leads to slightly lower results. However, when "Nightime" has to be detected, the retina is more efficient. It still eliminates the mean luminance, but here, it maximizes the signal to noise ratio, which would otherwise be low because of physical limitations of image sensors. Therefore, details are better extracted by the retina for the "Nighttime" concept, particularly the highly detailed and contrasted areas generated by the lights (e.g. streetlights, car headlights) of the visual scene. This shows that the retina generally has a good effect in the concept detection performance, however, the performance increase (or decrease) also depends on the retina's parameters. A more extended study will allow performances for more concepts to be improved by fine-tuning the retina's parameters for those concepts individually.

C. Computational cost

From a computational cost point of view, calculating the retinal outputs adds 18 products per pixel. The tradeoff between the added computational cost and the descriptor

enhancement obtained has to be considered from a global application level point of view, especially in the case of fusion-based approaches, where we need to compute several descriptors.

V. Conclusions

We proposed to enhance the performance of a classical SURF descriptor used by the state-of-the art Bag of Words approaches for High Level Features detection. The idea is to help such descriptors to be robust against spatio-temporal disturbances and to enhance the relevant information of the visual scene to process. This study shows that in the realistic and difficult case of the TrecVid challenge, this approach gives interesting results. It shows that the detection of HLFs from video keyframes can be significantly enhanced by preprocessing such keyframes with low-level human vision filtering (the parvocellular retina channel). The involved spatio-temporal properties of this channel help the BoW to better describe the static visual scene.

This preliminary study encourages us to investigate these preprocessing solutions further. We are particularly interested by their extension to other state-of-the-art descriptors (such as SIFT), to identify the performance boost that can be obtained. We expect descriptors that have a sensitivity to luminance changes, incorrect colors, image noise, compression artifacts or other image defects, to be particularly helped by the parvocellular processing.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA*, pages 1027–1035, 2007.

[2] N. Ballas, B. Delezoide, and F. Prêteux. Trajectories based descriptor for dynamic events annotation. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 13–18, New York, NY, USA, 2011. ACM.

[3] A. Benoit, A. Caplier, B. Durette, and J. Herault. Using human visual system modeling for bio-inspired low level image processing. *Computer Vision and Image Understanding*, 114(7):758 – 773, 2010.

[4] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University, 2009.

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

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV (2)'06*, pages 428–441, 2006.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40:5:1–5:60, May 2008.

[8] D. Gorisse, F. Precioso, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, R. Vieux, B. Mansencal, J. Benois-Pineau, H. Boujut, C. Morand, H. Jégou, S. Ayache, B. Safadi, Y. Tong, F. Thollard, G. Quénot, M., M. Cord, A. Benoit, and P. Lambert. IRIM at TRECVID 2010: Semantic Indexing and Instance Search. In *TREC online proceedings*, pages –, Gaithersburg, États-Unis, Nov. 2010. GDR ISIS.

[9] J. Hérault. *Vision: Signals, Images and Neural Networks*. Progress in Neural Processing. World Scientific Publishers, 2009. Département Images et Signal.

[10] L. Juan and O. Gwun. A comparison of sift , pca-sift and surf. *International Journal of Image Processing IJIP*, 3(4):143–152, 2009.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

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

[13] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2:215–322, April 2009.

[14] K. E. van de Sande, T. Gevers, and C. G. Snoek. A comparison of color features for visual concept classification. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 141–150, New York, NY, USA, 2008. ACM.

[15] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[16] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 102–111, New York, NY, USA, 2006. ACM.

[17] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 603–610, New York, NY, USA, 2008. ACM.