RETINA ENHANCED BAG OF WORDS DESCRIPTORS FOR VIDEO CLASSIFICATION

Sabin Tiberius Strat[*], *Alexandre Benoit*[†], *Patrick Lambert*[†]

[*] LISTIC - Université de Savoie
Annecy Le Vieux, France
LAPI - University "Politehnica" of Bucharest
Bucharest, Romania

[†] LISTIC - Université de Savoie
Annecy Le Vieux, France

ABSTRACT

This paper addresses the task of detecting diverse semantic concepts in videos. Within this context, the Bag Of Visual Words (BoW) model, inherited from sampled video keyframes analysis, is among the most popular methods. However, in the case of image sequences, this model faces new difficulties such as the added motion information, the extra computational cost and the increased variability of content and concepts to handle. Considering this spatio-temporal context, we propose to extend the BoW model by introducing video preprocessing strategies with the help of a retina model, before extracting BoW descriptors. This preprocessing increases the robustness of local features to disturbances such as noise and lighting variations. Additionally, the retina model is used to detect potentially salient areas and to construct spatio-temporal descriptors. We experiment with three state of the art local features, SIFT, SURF and FREAK, and we evaluate our results on the TRECVid 2012 Semantic Indexing (SIN) challenge.

Index Terms— Video, classification, retina, saliency, Bag of Words

1. INTRODUCTION

With the continuous and accelerated increase in size of video databases (on-line collections grow by thousands of hours each day), typical interactions such as indexing, browsing or searching require improved tools to organize the collection. These tools require an indexing of the collection, but due to the large amount of data, there is great interest in methods able to do this automatically. However, users require indexing with humanly-understandable terms, which is a challenging task due to the semantic gap: the transition from raw video data to high-level semantic concepts.

In the case of static image collections, one of the most popular methods for concept detection is the *Bag of visual Words* (BoW) [1] approach. It relies on extracting local visual features from the image, matching these features with a codebook of "visual words" and describing the image as a histogram of the visual words that were found. Supervised classification is used in the end to make the link between the BoW histogram and semantic concepts. Each part of this tool chain can be adjusted to improve classification performance regarding the application context, from image description to the final classification stage [2].

However, in the case of large video collections, because of computational cost, the video description stage of the BoW tool chain is very similar to the one for still images, thus omitting the added temporal (motion) information. Some spatio-temporal approaches have then been proposed, such as the MoSIFT local feature descriptor [3] or point trajectories [4], but these require significant computational resources. Improvements have also been made to take into account the temporal structure of actions, which the BoW model was lacking. The Actom Sequence Model of [5] was proven efficient in this respect, but it requires a precise annotation at the training stage of supervised classifiers.

In parallel, considering the high semantic level expected by end users, the introduction of visual perception has been suggested. Saliency maps can weigh the importance of collected visual words, which can improve tasks such as object recognition [6]. However, the complexity of saliency modeling is an important limitation in the analysis of large video collections. Other bio-inspired methods such as deep learning [7] have recently shown impressive results, but the underlying computational cost and architecture still limits their use.

All in all, when dealing with large video collections and unconstrained query topics, research challenges such as TRECVid [8] or MediaEval [9] show that spatio-temporal and multimodal content description is indeed required. Similarly, human perception follows the same multimodal philosophy. For example, the human retina enhances visual signals and decorrelates spatial and temporal information, which facilitates processing by specialized areas in the brain, where data is analysed, interpreted and fused at different stages.

This paper focuses on visual spatio-temporal content description by feeding a state of the art BoW approach with retina preprocessed information. We show that even though the retina model is a very low-level processing step in the visual system, it can significantly improve the results of a state

of the art BoW approach, it can help focus on potentially salient areas and it can process *spatio-temporal* information even with classical spatial local features such as SIFT, SURF or FREAK, while maintaining low computational cost.

We evaluate our approaches on the difficult Semantic Indexing (SIN) task of the TRECVid 2012 challenge, which requires detecting 346 diverse semantic concepts in a multitude of short video fragments (*video shots*). The rest of the paper is organized as follows: Sec. 2 describes our improved BoW toolchain using retinal preprocessing, Sec. 3 describes our experiments and Sec. 4 concludes the paper.

2. MODIFIED BOW TOOLCHAIN

2.1. The human retina model

As described in [10], we preprocess the input video stream with the human retina model of [11] (available in OpenCV), before applying the state of the art BoW toolchain. This retina model is able to remove spatio-temporal high frequency noise and whiten the image spectrum, thus providing enhanced signals for the following processing stages. The retina decorrelates spatial and temporal information by providing two output video streams:

- the *parvocellular* channel (parvo) (Fig. 1b and 1f) transmits static color details with reduced spatio-temporal high frequency noise. Contrast gain control ensures local adaptation to luminance. Spectral whitening also attenuates the mean luminance energy and enforces medium-frequency spatial components (details).
- the *magnocellular* channel (magno) (Fig. 1c and 1g) also benefits from local contrast adaptation and noise removal. It transmits low-resolution version of transient (moving) elements.

Based on the magnocellular channel, a low-level detector of spatio-temporal saliency can highlight areas where potentially more interesting content is located [10]. This detector has two phases: a transient phase (when processing starts, lasting about 5 frames) and a stable state. During the transient phase, because the retina reacts in a coarse to fine way, large texture rich areas of high luminance are first selected, as in Fig. 1d, where static faces from the background are detected along with the presenter's face. After a few frames, the response stabilizes and only transient (moving) areas are selected, such as the presenter's face and hands in Fig. 1h. The selected "blobs" are stable from one image to the next, which enforces the weight of features collected from such areas in the BoW histogram.

2.2. Feature collection strategies

Our main contribution relies in preprocessing the input videos to improve the local feature extraction step. Each of these local feature extraction strategies gives rise to its own visual vocabulary and BoW histogram. We use 5 such strategies, described in the following. All of them collect local features from a regular dense grid, with the same grid step in all cases. The first two are keyframe based approaches, extracting features from a single frame of the video shot. The other 3 consider a temporal window of 20 frames centered on the keyframe, and features are sampled only from grid points that fall within spatio-temporally "salient" regions defined by a binary mask (Fig. 1d and 1h).

The *baseline* consists in extracting local features in an opponent color space on the original unprocessed keyframe of the video shot. This approach serves as a reference for evaluating our other 4 strategies.

The *retina* strategy collects local features in an opponent color space from the parvocellular-preprocessed keyframe instead of the original keyframe. The goal is to take advantage of the "cleaner" parvocellular output to collect higher quality features than the baseline.

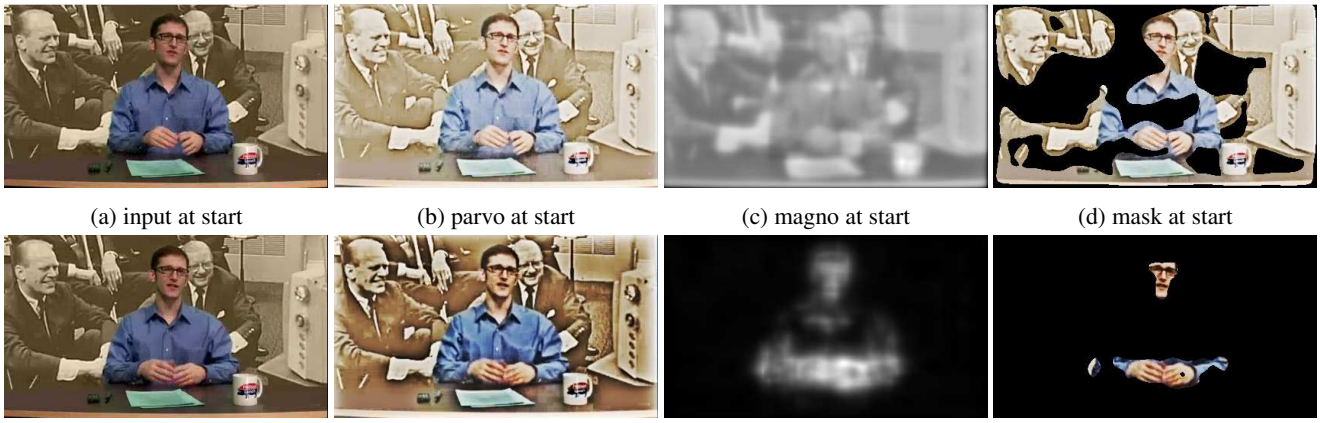
Retina masking parvo expands the *retina* strategy by collecting features in a *temporal window* of 20 frames centered on the keyframe. Additionally, the low-level saliency detector is used to select features that are hopefully more relevant. During the transient phase of the detector, large spatial features are selected, giving information about the general composition of the scene, while during the stable state, the contribution of moving features is enforced.

Retina masking magno collects local features from exactly the same positions and moments as *retina masking parvo*, but uses grayscale features from the magnocellular channel instead of color features from the parvocellular one. During the transient phase of the retina, these magno features constitute low-resolution data about the general appearance of the image (Fig. 1c). During the stable state, these features give a coarse description of contours perpendicular to the motion direction (Fig. 1g).

The last strategy, *multichannel masking*, also collects features at the same locations and moments as the previous two approaches. However, it employs hybrid local features by concatenating the SIFT/SURF/FREAK (opponent color space) signature of the image patch from the parvocellular channel with the SIFT/SURF/FREAK (grayscale) signature at the same location but from the magnocellular channel. During the stable state of the retina, the signature from the parvo channel encodes spatial appearance, while the signature from the magno channel encodes the motion direction. This makes the hybrid local features *spatio-temporal*.

2.3. Integration in the BoW framework

After preprocessing the video shots with the human retina model, we collect local features according to the strategies described above. Feature points are sampled on a dense grid with a 6 pixel step. 8 scales are used for image patch sizes, starting from 16x16 pixels, with an increasing scaling factor



(a) input at start (b) parvo at start (c) magno at start (d) mask at start

(e) input at end (f) parvo at end (g) magno at end (h) mask at end

Fig. 1: Effects of parvocellular and magnocellular preprocessing, as well as segmented "blobs" of low-level spatio-temporal saliency, at the beginning and end of a 20-frame temporal window around the keyframe. The retina is initialized at the beginning of this window.

of 1.2 between scales.

The rest of the processing chain is the same as for classical BoW approaches. The collected image patches (local features) are described with an image patch descriptor: SIFT [12] and SURF [12] are based on local histograms of image intensity gradients, while the more recent FREAK [13] constructs a binary signature by comparing pixel intensities in a retina-like pattern. All these image patch descriptors have some robustness to luminance changes, but they can be affected by noise, which is especially present in darker regions. Fortunately, the parvocellular channel benefits from reduced noise and improved local contrast.

Afterwards, for each of the 5 feature collection strategies and for each of the 3 local image patch descriptors, a specific visual vocabulary is computed, leading to 15 different vocabularies. Each one is extracted from the best kmeans clustering of 3 trials performed over 4 million video shot keypoints extracted from a training part of the dataset. With 1024 visual words, BoW histograms remain compact thus allowing fast extraction and classification. Finally, video shots are described by 15 histograms (the Bag of Words) depicting the frequencies of appearance of visual words. Histogram computation is performed using the soft-assignment strategy described in [14], with the specific $beta$ parameter equal to 10, which was proven to give better results than hard assignment (assigning only to the single closest vocabulary word).

The next step is to train and apply a supervised classifier that will give, for each semantic concept and each image/video, the likelihood of the concept being present in the multimedia element. We use the KNN classifier of [15], which is much faster to compute than SVM classifiers, even though the latter give better results. We prefer KNN because it is sufficient for comparing different video descriptors and it allows conducting more experiments in a shorter time. The interested reader can refer to [15] for a comparison of KNN and SVM classifiers, as well as for the effect of some descriptor optimisation strategies.

In the end, information fusion strategies can be used to take advantage of the entire set of descriptors that we propose.

3. EXPERIMENTS

3.1. Experimental setup

We perform our experiments on the TRECVid 2012 Semantic Indexing (SIN) development dataset. It contains cca. 200 hours of video divided into cca. 400 000 shots, in which the presence or absence of 346 diverse semantic concepts needs to be annotated [8]. The dataset is split in two parts, arbitrarily called *2012x* and *2012y*, each containing cca. 100 hours of video in cca. 200 000 shots. The two parts are balanced to contain similar numbers of true positives for each target concept. One of these parts is used to extract visual vocabularies and to train supervised classifiers, while the other part is used to evaluate results. For the evaluation dataset part, for each concept, a list of max. 2000 shots is to be returned, ranked according to decreasing likelihoods of containing the concept. The *inferred average precision* (infAP) [16], which is the official performance metric of TRECVid SIN, is used to evaluate the quality of these lists.

3.2. Individual results

Table 1 details the infAP obtained, averaged over all 346 concepts. Two results are given, one for training on *2012x* and testing on *2012y*, and the other for training on *2012y* and testing on *2012x*. These two results are similar, proving that the two datasets are balanced and that descriptor behaviours are stable. The performances might appear low, but these values

Individual descriptors	**SIFT**		**SURF**		**FREAK**	
infAp (%) [16]	xy	yx	xy	yx	xy	yx
baseline	8.7	8.5	4.4	4.1	8.3	8.3
retina	9.6	9.9	6.5	6.6	8.6	8.4
retina mask. parvo (P)	9.7	9.7	8.6	8.8	8.9	8.9
retina mask. magno (M)	7.7	7.6	7.1	7.0	4.7	4.8
multichannel mask.	9.0	9.5	8.9	8.7	8.4	8.3
Descriptor fusions						
retina mask. P+M early	9.8	9.8	8.9	9.1	8.8	8.8
F1:retina mask. P+M late	10	11	9.9	9.9	9.4	9.3
F2:F1+retina	12	12	11	11	11	11
F3:F1+retina+baseline	13	13	11	11	12	12
F3 SIFT+SURF+FREAK	15(yx)		15(xy)			

Table 1: Infered average precisions averaged over the 346 concepts. For each local feature type and each method we train on 2012x and report results on 2012y ('xy') and vice-versa ('yx') .

are normal considering the reduced number of true positives and the difficulty of the task [15]. As shown in [15], these results can be significantly improved with the aid of information fusion strategies that exploit the complementarity among descriptors. Even a simple arithmetic mean of classification scores coming from different descriptors already provides a significant boost.

Table 1 allows us to evaluate the gains obtained with the various proposed approaches compared to the *baseline*. The two strategies employing features collected on the parvocellular frame(s), namely *retina* and *retina masking parvo*, improve results for all 3 types of local features. The relative gain is the greatest for SURF features, where performance is doubled compared to the baseline. The baseline for SURF is the lowest, but the retina approaches manage to bring it to a similar level as SIFT of FREAK.

A large part of this gain is due to the image cleaning and detail enhancement effect provided by the parvocellular channel. It can be directly observed when comparing the 2 keyframe based approaches, *baseline* and *retina* whatever the local features used. The SURF based BoW has the lowest baseline and is improved by 48%, while SIFT and FREAK are respectively improved by 10% and 4%.

Also, we defend the importance of focusing on salient areas within the considered video shots. Comparing the keyframe-based *retina* and the temporal window with saliency masking *retina masking parvo*, no improvement is observed for SIFT which already has high performance with the *retina* description. However improvements amount to 3-6% for FREAK and 32% for SURF. In the end, with the *retina masking parvo* strategy, the 3 different local features SIFT/SURF/FREAK get similar results close to 9% infAP.

With the exception of SURF, the *retina masking magno* approach, which collects features from the magnocellular channel, gives lower results than the baseline on average over the 346 concepts. Indeed, since this channel only contains low spatial frequencies of temporally-transient signals, this channel provides a coarse representation that should be described only at high scales. Nevertheless, compared to its *parvo* channel counterpart, it improves performance for some specific concepts: with SIFT, 74 concepts such as "First Lady", "HighWay", "Ski" benefit from the magnocellular channel; with SURF, 93 concepts such as "Skating", "Soccer Player" or "Indian Person"; and with FREAK, 25 concepts such as "Primate" and once again "Skating". A detailed concept per concept analysis shows that very few concepts are classified similarly with the 3 types of local features. This announces a complementarity between these 2D features that will be discussed later on.

3.3. Descriptor fusions

A first type of fusion has already been proposed in the form of the *multichannel masking* approach, which uses hybrid parvo-magno local features, first proposed in [10]. However, in our experiments, global results are similar to the *baseline*. Only SURF descriptors experience a doubling of performance compared to the *baseline*, but the *retina masking parvo* approach, using parvo features instead of parvo-magno, already gives almost the same result.

The alternative is to perform fusions at higher levels. An *early fusion* is obtained by concatenating the BoW histogram from *retina masking parvo* with the one from *retina masking magno* before the supervised classification step (this fusion is denoted *retina masking P+M early* in Table 1). It achieves only a 6% improvement compared to the *multichannel* strategy in the case of FREAK descriptors, and performances remain close to using the *retina masking parvo* approach alone.

Moving on to an even higher fusion level, we perform a *late fusion* of the *retina masking parvo* and *retina masking magno* approaches by averaging their supervised classification scores (denoted *F1: retina masking P+M late* in Table 1). This time, the performance increase is much more significant, of 12% compared to the *multichannel* strategy in the case of FREAK descriptors and similar amounts for SIFT and SURF. This late fusion also exceeds not only the *baseline*, but all the other individual BoW descriptors.

If we now enrich this late fusion scheme, we can exploit the complementarity of all the presented video description methods. If we combine the *retina* classification scores with the ones from *F1: retina masking P+M late*, obtaining the *F2: F1+retina* approach in Table 1, an additional increase of 18% is obtained compared to *F1*, in the case of SIFT descriptors. If we also include in this late fusion the *baseline* approach, obtaining *F3: F1+retina+baseline*, we have an additional 7% increase compared to *F2* with SIFT features. This shows that the keyframe and masking descriptors are comple-

mentary, and that even the baseline descriptor remains important because it also deals with high spatial frequencies that are otherwise cut by the parvocellular channel.

In the end, if we consider the complementarity between descriptors based on SIFT, SURF and FREAK local features, fusing the *F3* methods allows performance to reach its maximum. Compared to the SIFT *baseline* a gain of 66% is obtained. This simple fusion method opens perspective for enhanced results using more specialised fusion methods as the ones proposed in [15].

4. CONCLUSION

In this paper, we show that a retina model applied before a state of the art Bag of Words approach improves visual concept detection. Performance boosts have been observed on different types of image signatures from SIFT to binary FREAK. This retinal model also allows a set of very different and complementary descriptors to be designed, which can lead to even better results with the aid of even a simple late fusion approach. Future work will address an improved retina model, with less motion blur in the parvocellular channel but which retains the spatio-temporal low-level saliency property of the magnocellular channel.

Acknowledgements

Our descriptors have been generated using the MUST computing center of the University of Savoie. KNN classification and performance evaluations were carried out on the Grid'5000 platform, using the IRIM group (Indexation et Recherche d'Information Multimedia) software tools developed under the INRIA ALADDIN initiative with support from CNRS, RENATER and several universities as well as other funding bodies (see `https://www.grid5000.fr`).

REFERENCES

[1] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct., pp. 1470–1477 vol.2.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," 2011.

[3] M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," Tech. Rep. CMU-CS-09-161, Carnegie Mellon University, 2009.

[4] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *ICCV 2013 - IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013, IEEE.

[5] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom Sequence Models for Efficient Action Detection," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, 2011, MSR-INRIA.

[6] Iván González-Díaz, Hugo Boujut, Vincent Buso, Jenny Benois-Pineau, and Jean-Philippe Domenger, "Saliency-based object recognition in video," 10 pages, 2013.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *ArXiv e-prints*, June 2012.

[8] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[9] S. Little, A. Llorente, and S. Rger, "An overview of evaluation campaigns in multimedia retrieval," in *ImageCLEF*, vol. 32 of *The Information Retrieval Series*, pp. 507–525. Springer Berlin Heidelberg, 2010.

[10] S.T. Strat, A. Benoit, and P. Lambert, "Retina enhanced SIFT descriptors for video indexing," in *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, June 2013, pp. 201–206.

[11] 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*, vol. 114, no. 7, pp. 758 – 773, 2010.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[13] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 510–517.

[14] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2486–2493.

[15] N. Ballas et al, "IRIM at TRECVID 2013: Semantic Indexing and Instance Search," in *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVid)*, Gaithersburg, MD, USA, Nov. 2013.

[16] 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*, New York, NY, USA, 2008, SIGIR '08, pp. 603–610, ACM.