Learned features versus engineered features for semantic video indexing

Mateusz Budnik, Efrain Leonardo Gutierrez Gomez, Bahjat Safadi, Georges Quénot

Learned features versus engineered features for semantic video indexing

Mateusz Budnik[1,2] Efrain-Leonardo Gutierrez-Gomez[1,2] Bahjat Safadi[1,2] Georges Quénot[1,2]
[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]CNRS, LIG, F-38000 Grenoble, France
{Firstname.Lastname}@imag.fr

***Abstract*—In this paper, we compare "traditional" engineered (hand-crafted) features (or descriptors) and learned features for content-based semantic indexing of video documents. Learned (or semantic) features are obtained by training classifiers for other target concepts on other data. These classifiers are then applied to the current collection. The vector of classification scores is the new feature used for training a classifier for the current target concepts on the current collection. If the classifiers used on the other collection are of the Deep Convolutional Neural Network (DCNN) type, it is possible to use as a new feature not only the score values provided by the last layer but also the intermediate values corresponding to the output of all the hidden layers. We made an extensive comparison of the performance of such features with traditional engineered ones as well as with combinations of them. The comparison was made in the context of the TRECVid semantic indexing task. Our results confirm those obtained for still images: features learned from other training data generally outperform engineered features for concept recognition. Additionally, we found that directly training SVM classifiers using these features does significantly better than partially retraining the DCNN for adapting it to the new data. We also found that, even though the learned features performed better that the engineered ones, the fusion of both of them perform significantly better, indicating that engineered features are still useful, at least in this case.**

I. Introduction

Deep Convolutional Neural Networks (DCNN) have recently made a significant breakthrough in image classification [1]. This has been made possible by a conjunction of factors including: findings about how to have deep networks effectively and efficiently converge [2], the use of convolutional layers [3][4], the availability of very powerful parallel architectures (GPUs), findings about how exactly a network should be organized for the task [1], and the availability of huge quantity of cleanly annotated data [5].

Not to minimize the importance of the hardware progress and of algorithmic breakthroughs, the availability of a large number of image examples for a very large number of concepts was really crucial as DCNNs really needs such amount of training data for actually being efficient. Data augmentation (e.g. multiple crops of training samples) can further help but also only when a huge amount of data is already available. Such amount of training data is currently available only with ImageNet which corresponds to a single type of application and only for still images. For video documents for instance, many annotated collection exist but with much smaller number of concepts and/or much smaller number of examples. Trying to train DCNNs on such data generally leads to results that are less good than those obtained using "classical" engineered features (or descriptors) combined with also "classical" machine learning methods (typically SVMs).

Two strategies have been considered for making other domains benefit from the success of the DCNN/ImageNet combination. The first one consists in pre-training a DCNN on ImageNet data and then partly retrain or fine-tune it on another collection [6][7]. Generally, only the last layers are retrained, the exact number of which as well as the learning parameters being experimentally determined by cross-validation. Though this strategy can produce much better results than when the DCNN is trained only on the target data, it does not necessarily compete with classical approaches and/or lead to gains that are much less important than in the ImageNet case.

The second strategy consists in using a DCNN pre-trained on ImageNet, applying it to a different target collection and use the final ImageNet concept detection scores and/or the output of the hidden layers as features for training classifiers and making prediction on the different target collection. Razavian et al. [8] have very successfully applied this strategy to a number of test collections for both image indexing and image retrieval.

In this work, we explore how these strategies perform in the context of video indexing. We also investigate how they can be combined with classical methods based on engineered features and how they can be combined with other video-specific improvement methods like temporal re-scoring [9]. Experiments have been carried out in the context of the semantic indexing task at TRECVid [10]. In this paper we make the following contributions:

1) We confirm the results obtained for still images in the case of video shot indexing: features learned from other training data generally outperform engineered features for concept recognition.
2) We show that directly training SVM classifiers using these features does better than partially retraining the DCNN for adapting it to the new data.
3) We show that, even though learned features outperform engineered ones, fusing them perform significantly better, indicating that engineered features are still useful, at least in this case.
4) We show that temporal an conceptual re-scoring methods also improve classification results obtained with DCNN features.

II. Related work

Semantic features are not restricted to DCNN and had already been used for multimedia classification and retrieval.

Smith et al. [11] introduced them as "model vectors". These provide a semantic signature for multimedia documents by capturing the detection of concepts across a lexicon using a set of independent binary classifiers. Ayache et al. [12] proposed to use local visual categories detection scores on regular grids or to use topic detection on ASR transcriptions for video shot classification. Su et al. [13] also proposed to use semantic attributes obtained with supervised learning either as local or global features for image classification.

In all these works and many other similar ones, the semantic features are learned on completely different collections and generally for concepts or categories different from those searched for on the target collection. Hamadi et al. [14] used the approach using the same collection and the same concepts both for the semantic feature training and for their use in a further classification step. In this variant, called "conceptual feedback", a given target concept is learned both from the "low-level" features and from the detection scores of the other target concepts also learned from the same low-level features (the training of the semantic features has to be done by cross-validation within the training set so that it can be used for the second training step both on the training and test sets).

Concerning the first DCNN transfer strategy (DCNN re-training), Yosinski et al. [7] et al showed that the features corresponding to the output of the hidden layers are well transferable from one collection to another and that re-training only the last layers is very efficient both for comparable or for dissimilar concept types. Their experiments were conducted only within the ImageNet collection however. Similar results were obtained by Chatfield et al. [6] on different data.

Concerning the second DCNN transfer strategy (classical training with features produced by DCNNs), Razavian et al. [8] showed that it works very well too, for several test collection, some of which are close to ImageNet and some of which are quite different both in terms of visual contents and in terms of target concepts. They also showed that this type of semantic features can be successfully used both for categorization tasks and for retrieval tasks. Finally, they showed that in addition to the score values produced by the last layer, the values corresponding to the output of all the hidden layers can be used as feature vectors. The semantic level of the layers output values increases with the layer number from low-level, close to classical engineered features for the first layers, to fully semantic for the last layers. Their experiments showed that using the last but one and last but two layer outputs generally gives the best results. This is likely because the last layers contain more semantic information while the last one has lost some useful information as it is tuned to different target concepts. There is generally no equivalent to the output of the hidden layers in classical learning methods (e.g. SVMs) and these can only produce the final detection scores as semantic features.

Many variants of the "classical" approach exist. Most of them consist in a feature extraction step followed by a classification step. In many cases, several different features can be extracted and in some cases a few different classification methods are also used in parallel; a fusion step has then to be considered. Fusion is called "early" when it is performed on extracted features, "late" when it is performed on classification scores or "kernel" when it is performed on computed kernel within the classification step (for kernel-based methods); many combinations can also be considered.

A very large variety of engineered features has been designed and used for multimedia classification. Some of them are directly computed on the whole image (e.g. color histograms), some of them are computed on image parts (e.g. SIFTs) [15]. In the latter case, the locally extracted features need to be aggregated in order to produce a single fixed-size global feature. Many methods can be used for that, including the "bag of visual words" one (BoW) [16][17] or the Fisher vector (FV) one [18] and similar ones like super vectors (SV), and vector of locally aggregated descriptors (VLAD) [19] or tensors (VLAT) [20]. Some of them may reach their maximum efficiency only when they are highly dimensional, typically the FV, VLAD and VLAT ones. Two different strategies can be considered for dealing with them: either use linear classifiers combined with compression techniques [18] or using dimensionality reduction techniques combined with non-linear classifiers [21]. In the case of video indexing, engineered features have been proposed also for the representation of audio and motion content.

The comparison of methods presented here has been conducted in the context of the Semantic INdexing (SIN) task TRECVid [10]. It differs from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in many respects. Indeed, the indexed units are video shots instead of still images. The quality and resolution of the videos are quite low (512 and 64 kbit/s for the image and audio streams respectively). The target concepts are different: 346 non-exclusive concepts with generic-specific relations. Many of them are very infrequent in both the training and test data. The way the collection has been built is also very different. In ImageNet, a given number of sample images have been selected and checked for each target concept resulting in a high quality and comparable example set size for all concepts. In TRECVid SIN, videos have been randomly selected from the Internet Archive completely independently of the target concepts; the target concepts have been annotated a posteriori resulting in very variable number of positive and negative examples for the different concepts. Most of the concepts are very infrequent and also not very well visible. Compared to ImageNet, the positive samples are much less typical, much less centered, of smaller size and with a much lower image quality. The task is therefore much more difficult than the ILSVRC one but it may also be more representative of indexing and retrieval tasks "in the wild". An active learning method was used for driving the annotation process for trying to reduce the imbalance class effect in the training data and also ensure a minimum number of positive samples for each target concept [22]. The resulting annotation is sparse (about 15% in average) and consists in 28,864,844 concept $\times$ shots judgements. All of these differences probably explain why training DCNNs directly on TRECVid SIN data gives much poorer results than on ImageNet data and why the two considered adaptation strategies are needed (or perform much better) in this case.

III. Comparison between engineered features and semantic features

In this section, we compare the performance of engineered features and semantic features. For the engineered features, we

use a series of features shared by the participants of the IRIM group of the French GDR ISIS [23].

Two different classifiers have been used, one based on the k nearest neighbors and one based on Multiple SVMs because it handles well large class imbalances [9]. The predictions of these two classifiers were fused producing a globally better result [24]. We also use the approach based on dimensionality reduction combined with non-linear classification for dealing with high-dimensional features.

We considered the following engineered features types:

- CEALIST/bov_dsiftSC_8192: bag of visual terms [25]. Dense SIFT are extracted every 6 pixels. The codebook of size 1024 is built with K-means. Bags are generated with soft coding and max pooling. The final signature results from a three-level spatial pyramid: $1024 \times (1 + 2 \times 2 + 3 \times 1) = 8192$ dimensions.
- CEALIST/bov_dsiftSC_21504: bag of visual terms [25]. Same as CEALIST/bov_dsiftSC_8192 with a different spatial pyramid: $1024 \times (1 + 2 \times 2 + 4 \times 4) = 21504$ dimensions.
- ETIS/global_<feature>[<type>]x<size>: (concatenated) histogram features[26], where:
  - <feature> is chosen among lab (CIE L*a*b* colors) and qw (quaternionic wavelets, 3 scales, 3 orientations)
  - <type> can be m1x1 (histogram computed on the whole image), m1x3 (histogram for 3 vertical parts) or m2x2 (histogram on 4 image parts)
  - <size> is the dictionary size, sometimes different from the final feature vector dimension.

  For instance, with <type>=m1x3 and <size>=32, the final feature vector has 3 × 32 = 96 dimensions.
- ETIS/vlat_<desc type>_dict<dict size>_<size>: compact Vectors of Locally Aggregated Tensors (VLAT [20]). <desc type> = low-level descriptors, for instance hog6s8 = dense histograms of gradient every 6 pixels, 8×8 pixels cells. <dict size> = size of the low-level descriptors dictionary. <size> = size of feature for one frame. Note: these features can be truncated. These features must be normalized to be efficient (e.g. L_2 unit length).
- LIG/opp_sift_<method>[_unc]_1000: bag of word, opponent sift, generated using Koen Van de Sande's software[27]: 1000 dimensions (384 dimensions per detected point before clustering; clustering on 535117 points coming from 1000 randomly chosen images). <**method**> method is related to the way by which SIFT points are selected: **har** corresponds to a filtering via a Harris-Laplace detector and **dense** corresponds to a dense sampling; the versions with **_unc** correspond to the same with fuzziness introduced in the histogram computation.
- LIRIS/OCLPB_DS_4096: Dense sampling OCLBP [28] bag-of-words descriptor with 4096 k-means clusters. We extract orthogonal combination of local binary pattern (OCLBP) to reduce original LBP histogram size and at the same time preserve information on all neighboring pixels. Instead of encoding local patterns on 8 neighbors, encoding is performed on two sets of 4 orthogonal neighbors, resulting in two independent codes. Concatenating and accumulating two codes leads to a final 32 dimensional LBP histogram, compared with original 256 dimensions. 4096-dimensional bag-of-words descriptors are finally generated using a pre-trained dictionary.
- LISTIC/SIFT_*: Bio-inspired retinal preprocessing strategies is applied before extracting Bag of Words of Opponent SIFT features (details in [29]) using the retinal model from [30]. Features extracted on dense grids on 8 scales (initial sampling=6 pixels, initial patch=16x16pixels), using a linear scale factor 1.2. K-means clustering is used for producing dictionaries of 1024 or 2048 visual words. The proposed descriptors are similar to those from [29] except that multi-scale dense grids are used. Despite showing equivalent mean average performance, the various pre-filtering strategies present different complementary behaviors that boost performances at the fusion stage [31].

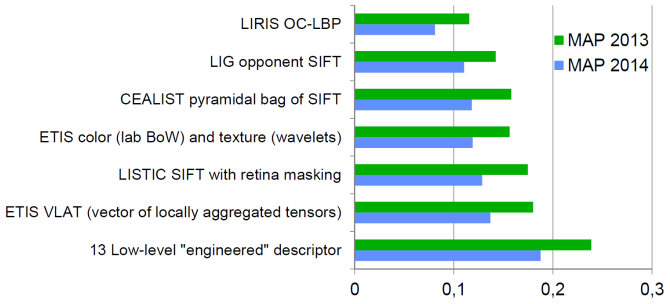

Fig. 1. Performance of engineered features

Figure 1 shows the performance of several types of engineered features. In several cases, the result is shown for already a combination of variants of the same feature type, for instance corresponding to a pyramidal image decomposition. The performance is given as the Mean (inferred) Average Precision on the 2013 and 2014 editions of the TRECVid SIN task. That task was a bit harder in 2014 because the set of evaluated concepts was different, including more difficult ones. The last entry in the figure correspond to a late fusion of all the IRIM features (including others less good than those shown). We can see that the fusion of these features does significantly better than the best of them.

We considered the following learned or semantic feature types:

- Semantic features computed using a Fisher vector approach [18]. Two variants were produced by Florent Perronnin fro Xerox: a 1000-dimensional one trained using the 1000 concepts of ILSVRC 1000, and a 10174-dimensional one trained using 10174 ImageNet concepts.
- Semantic features computed using a DCNN following the Krizhevsky architecture [1], using the caffe implementation [32], and pre-trained on the 1000 ISLVRC 2012 concepts. Two variants were produced by LIG

and Eurecom, corresponding respectively to the simple and to the data augmented versions.

- Quasi-semantic features corresponding to last three hidden layers of the same network (simple version) whose dimensionalities are respectively 43,264, 4,096 and 4,096 for layers 5, 6 and 7.
- Concepts features corresponding to the conceptual feedback approach [14] applied two times. These are were originally designed for being used with engineered descriptors but they acn also include other semantic descriptors; here they have been computed including the Xerox semantic descriptors.

Several late fusions of features of the same type were also considered.

Figure 2 shows the performance of several types of learned features. The first entry is the combination of all engineered features showed as a baseline. We can immediately see that almost all semantic features perform similarly to or better than the baseline and therefore significantly better than any individual engineered feature. We can also observe than combinations of semantic features perform even better. Considering the caffe output and internal layers, the best choice is fc6 which is very close to fc7. The final output layer is less good and fc5 is even less good. It is interesting to notice that the performance of the Xerox Fisher vector based semantic features is very close to the performance of the final output of the pre-trained caffe network while both use very similar training data (ILSVRC10 and ILSVRC12 respectively). The features corresponding to the conceptual feedback perform better but the Xerox semantic features were included in their production.

IV. Comparison between partial DCNN retraining and use of DCNN layer output as features

We made several trials for retraining the last layers of the pre-trained caffe implementation. We tried to retrain the last one, last two or last three layers, each time doing our best to select the optimal training parameters in each case. The best performance was obtained when retraining the last two layers and it was of 0.2171 and 0.1839 respectively on SIN 2013 and 2014 which is less that using the fc6 layer for which the KNN/MSVM classifier combination gives 0.2347 and 0.2016 respectively. While the retraining of the last two layers starts from the same fc6 feature, it seems that these two layers are not able to do a learning as good as the KNN/MSVM combination. This may be because they actually implement only a two-layer perceptron and because they have difficulties with highly imbalanced training data.

V. Temporal re-scoring with semantic features

We applied the temporal re-scoring method proposed by Safadi et al. [9] as it is a simple way to obtain a significant performance boost at very low computing cost. It simply exploits the fact that if a concept appears in a video shot, it is more likely to appear in the preceding and following shots.

Figure 3 shows the performance gain brought by the temporal re-scoring for the different considered semantic features on SIN 2014. We can observe that the gain is more important for some features that for others. It is greater for learned features than for engineered ones. It is greater for DCNN output features than for Fisher vector based features. It is also greater for semantic features than for quasi-semantic ones (corresponding to internal layers). It also appears smaller for conceptual feedback features but this is due to the fact that the feedback features were computed from scores that were already re-scored as it works better like this [14].

VI. Fusion and other improvement methods

Figure 4 shows the performance gain brought by the successive fusion of descriptors of increasing performance. This strategy has been selected as it has been observed that fusion gives better results when done on sources of comparable performance [33]. This strategy also ensures a dilution and therefore a smaller contribution of those having lower performances; it has also been observed that this is better than simply dropping them [33]. These fusions are followed by post-processing techniques bringing additional performance gains. The (D_LIG.14_X) labels correspond to our official submissions at the TRECVid 2014 SIN task.

Fusion of engineered and learned features does better than any of them separately, indicating that though they are individually and collectively less good, engineered features are still useful for global system performance. Though hidden layers are individually the best ones, they do not bring further improvement when the engineered descriptors have already been fused with Fisher vector based features and DCNN output features. Conceptual feedback and temporal re-scoring still improve further. Finally, two other improvement techniques were tried: conceptual re-scoring and use of an uploader model. Conceptual re-scoring is different from conceptual feedback, it is similar to temporal re-scoring but it exploits the semantic similarity between concepts instead of the temporal closeness between video shots [14]. It did not prove useful probably because, even if based on a different method, it captures the same type of information as the conceptual feedback done previously. The uploader model uses the information of who uploaded the video on the Internet archive and tries to exploit the fact that videos uploaded by a same user tend to have the same type of content [34]. It did not bring further improvement after all the other improvements were performed.

VII. Conclusion

In this paper, we have compared "traditional" engineered features and learned features for content-based semantic indexing of video documents. We made an extensive comparison of the performance of learned features with traditional engineered ones as well as with combinations of them. The comparison was made in the context of the TRECVid semantic indexing task. Our results confirm those obtained for still images: features learned from other training data generally outperform engineered features for concept recognition. Additionally, we found that directly training SVM classifiers using these features does better than partially retraining the DCNN for adapting it to the new data. We also found that, even though the learned features performed better that the engineered ones, the fusion of both of them still perform significantly better, indicating that engineered features are still useful, at least in this case.

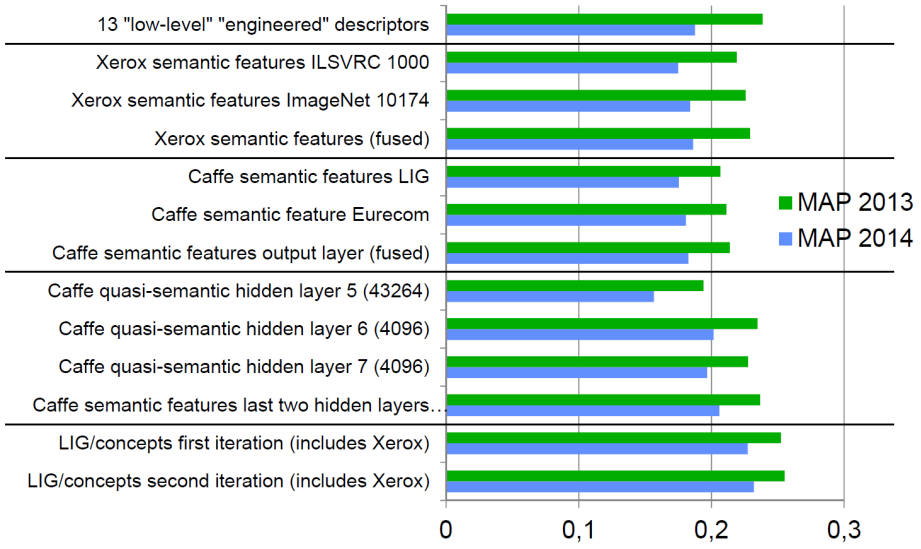

Fig. 2. Performance of semantic features

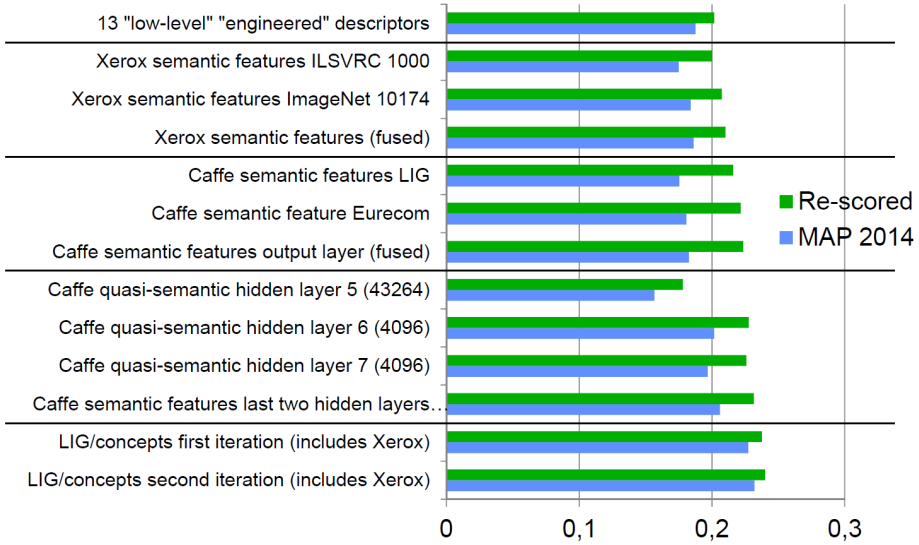

Fig. 3. Performance gain from temporal re-scoring with semantic features

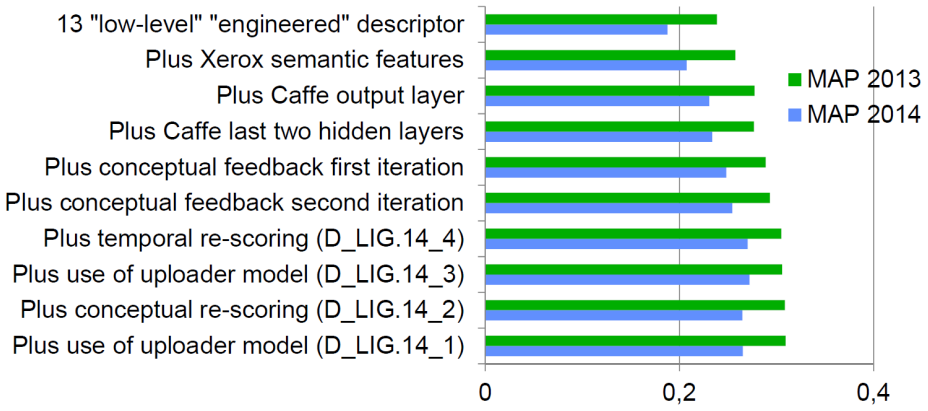

Fig. 4. Performance gain from fusion and other improvement methods

ACKNOWLEDGMENT

This work was conducted as a part of the CHIST-ERA CAMOMILE project, which was funded by the ANR (Agence Nationale de la Recherche, France). Part of the computations presented in this paper were performed using the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche. Results from the IRIM network were also used in these experiments [23]. The authors also wish to thank Florent Perronnin from XRCE for providing features based on classification scores from classifiers trained on ILSVRC/ImageNet data [18].

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] G. B. Orr and K.-R. Mueller, Eds., *Neural Networks : Tricks of the Trade*, ser. Lecture Notes in Computer Science. Springer, 1998, vol. 1524.

[3] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014.

[8] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2014, pp. 512–519.

[9] B. Safadi and G. Quénot, "Re-ranking by Local Re-scoring for Video Indexing and Retrieval," in *CIKM 2011 - International Conference on Information and Knowledge Management*, I. R. Craig Macdonald, Iadh Ounis, Ed. Glasgow, United Kingdom: ACM, Oct. 2011, pp. 2081–2084, poster session: information retrieval.

[10] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot, "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[11] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 2, July 2003, pp. II–445–8 vol.2.

[12] S. Ayache, G. Quénot, and J. Gensel, "Image and video indexing using networks of operators," *EURASIP Journal on Image and Video Processing*, vol. 2007, no. 1, p. 056928, 2007.

[13] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 59–77, 2012.

[14] A. Hamadi, P. Mulhem, and G. Quénot, "Extended conceptual feedback for semantic multimedia indexing," *Multimedia Tools and Applications*, vol. 74, no. 4, pp. 1225–1248, 2015.

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

[16] J. Sivic and 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*, ser. ICCV '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 1470–.

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

[18] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[19] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[20] D. Picard and P.-H. Gosselin, "Efficient image signatures and similarities using tensor products of local descriptors," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 680–687, Mar. 2013.

[21] B. Safadi, N. Derbas, and G. Quénot, "Descriptor optimization for multimedia indexing and retrieval," *Multimedia Tools and Applications*, vol. 74, no. 4, pp. 1267–1290, 2015.

[22] S. Ayache and G. Quénot, "Video Corpus Annotation using Active Learning," in *European Conference on Information Retrieval (ECIR)*, Glasgow, Scotland, mar 2008, pp. 187–198.

[23] N. Ballas, B. Labbé, H. Le Borgne, P. Gosselin, D. Picard, M. Redi, B. Mérialdo, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, N. Derbas, M. Budnik, G. Quénot, B. Gao, C. Zhu, Y. Tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benoit, P. Lambert, and T. Strat, "IRIM at TRECVID 2014: Semantic Indexing and Instance Search," in *Proceedings of TRECVID*, Orlando, United States, Nov. 2014.

[24] B. Safadi, N. Derbas, A. Hamadi, M. Budnik, P. Mulhem, and G. Quénot, "LIG at TRECVid 2014: Semantic Indexing," in *Proceedings of TRECVID*, Orlando, United States, Nov. 2014.

[25] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 3618–3625.

[26] P. H. Gosselin, M. Cord, and S. Philipp-Foliguet, "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 403 – 417, 2008, similarity Matching in Computer Vision and Multimedia.

[27] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[28] C. Zhu, C.-E. Bichot, and L. Chen, "Color orthogonal local binary patterns combination for image region description," *Rapport technique RR-LIRIS-2011-012, LIRIS UMR*, vol. 5205, p. 15, 2011.

[29] S. 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.

[30] 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.

[31] S. Strat, A. Benoit, and P. Lambert, "Retina enhanced bag of words descriptors for video classification," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, Sept 2014, pp. 1307–1311.

[32] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[33] S. Strat, A. Benoit, H. Bredin, G. Qunot, and P. Lambert, "Hierarchical late fusion for concept detection in videos," in *Computer Vision ECCV 2012. Workshops and Demonstrations*, A. Fusiello, V. Murino, and R. Cucchiara, Eds., 2012, pp. 335–344.

[34] U. Niaz and B. Merialdo, "Improving video concept detection using uploader model," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.