EURECOM at TrecVid 2015: Semantic Indexing and Video Hyperlinking Tasks

Usman Niaz, Bernard Merialdo, Claudiu Tanase, Maria Eskevich, Benoit Huet
Multimedia Department, EURECOM
Sophia Antipolis, France
firstname.surname@eurecom.fr

October 26, 2015

1 Abstract

This year EURECOM participated in the TRECVID 2015 Semantic INdexing (SIN) Task [24] for the submission of four different runs for 60 concepts, and Video Hyperlinking (LNK) Task [24] with the four submissions. Our submission to the SIN Task builds on the runs submitted in the previous years for the 2013 and 2014 SIN tasks, the details of which can be found in [20] and [19], while the LNK submissions are based on our previous experiments as in [28] and [9].

The major changes for 2015 are the use of new Deep Network models to produce extra descriptors for the video shots, and the introduction of various fusion schemes at all levels of the processing, to reduce the problem of overfitting. This year, we did not use our uploader model, partly because of lack of time, and partly because initial experiments showed only marginal improvement after the new features were added.

For the LNK Task our approach targeted to connect the textual stream of the videos within the collection and its vocabulary context, as defined by word2vec algorithm, with the output of visual concepts detection tools for the corresponding hyperlinks candidates within one framework. We combined visual concepts detection confidence scores with the information about corresponding word vectors distances in order to rerank the baseline text based search. The reranked runs did not outperform the baseline, however they exposed potential of our method for further improvement.

Beside this participation, EURECOM took part in the collaborative IRIM submission, the details of this contribution is included in the corresponding publication from the IRIM group.

The remainder of this paper briefly describes the descriptors that we have been using, the training and the various fusion schemes, and the content of the submitted runs; and the framework of the confidence scores combinations used for reranking in the LNK task.

2 EURECOM Basic Descriptors for SIN

We are using the following set of engineered visual features ranging from local features to global image descriptions.

- **Color Moments** This global descriptor computes, for each color channel in the LAB space, the first, second and third moment statistics on 25 non overlapping local windows per image.
- **Wavelet Feature** This texture-based descriptor calculates the variance in the Haar wavelet sub-bands for each window resulting from a 3×3 division of a given keyframe.
- **Edge Histogram** The MPEG-7 edge histogram describes the edges' spatial distribution for 16 sub-regions in the image.
- **Local Binary Pattern (LBP)** Local binary pattern describes the local texture information around each point [21], which has been proven effective in object recognition. We employ the implementation in [2] to extract and combine the LBP features with three different radius (1, 2, and 3) and get a 54-bin feature vector.
- **SIFT from keypoints** Two sets of interest points are identified using different detectors:

  1. Difference of Gaussian
  2. Hessian-Laplacian Detector

  For each of the detected keypoints we then compute a SIFT [16] descriptor using the VIREO system [3]. We use the K-means algorithm to cluster the descriptors from the training set into 500, 1,000 and 2,000 visual words. After quantization of the feature space, an image is represented by a histogram where the bins of this histogram count the visual words closest to image keypoints. We therefore obtain feature vectors of dimension 5,00, 1,000 and 2,000. This run uses only the 2,000 vocabulary.
- **Dense SIFT and ColorSIFT** We also use a dense sampling for the SIFT and ColorSIFT descriptors proposed by Koen Van de Sande [33]. We use their software provided in [1]. We created visual dictionaries of size 1,000, 4,000 and 10,000. We pool the quantized descriptors globally over the whole image. We also consider pooling according to a spatial pyramid (1, 2x2, 3x1), so that the corresponding feature vectors have a dimension 8 times the size of the dictionary. This run uses only the 10,000 vocabulary.
- **Saliency Moments descriptor** This is a holistic descriptor which embeds some locally-parsed information, namely the shape of the salient region, in a holistic representation of the scene, structurally similar to [22]. First, the saliency information is extracted at different resolutions using a spectral, light-weight algorithm [12]. The signals obtained are then sampled directly in the frequency domain, using a set of Gabor wavelets. Each of these samples, called "Saliency Components", is then interpreted as a probability distribution:

the components are divided into subwindows and the first three moments are extracted, namely mean, standard deviation and skewness. The resulting signature vector is a 482-dimensional descriptor [26].

- **MEDA** We have proposed descriptors based on marginal distributions of the local descriptors. They have the advantage of a faster computation than bag-of-word construction, and have shown efficient performance. Those descriptors are described in [27].
- **ST-MPEG7** This is a spatio-temporal descriptor based on the temporal statistics of the MPEG-7 Edge Histogram descriptor.

3 EURECOM Deep Networks Descriptors for SIN

We are also using features extracted using Deep Networks. As the training of these networks is quite computer intensive, at this stage we just rely on the use of existing pre-trained networks. We have used the following networks, which have been trained on the ImageNet corpus.

- **Caffe AlexNet** This is one of the models available from the Caffe framework [13]. It uses the architecture of the network described in [14] and is trained on the ILSVRC 2012 task of ImageNet. We apply this network directly over the development and test sets of TRECVid. The output of the network is a 1,000 values vector which is used as a descriptor for each keyframe. We also use the values of the last hidden layer, which is a 4,096 values vector. This makes our **caffe1000** and **caffe4096** descriptors.
- **VGG Very Deep Networks** We use the Very Deep Networks made available by the Oxford Visual Geometry Group [31]. We use both the 16 layer and 19 layer models. Both models have been trained on ILSVRC 2012 task of ImageNet. We apply these networks directly over the development and test sets of TRECVid. From each model, we extract three descriptors:
  - the output of the network, which is a 1,000 values vector corresponding to the concepts of ImageNet,
  - the output of the last hidden layer, which is a 4,096 values vector,
  - the output of the second to last hidden layer, which is again a 4,096 values vector.

  This provides a total of 6 descriptors from the VGG networks.

Thus we have a total of 8 descriptors which are extracted using Deep Networks.

4 EURECOM Runs for SIN

4.1 Training classifiers

All our runs use SVM classifiers that are trained on the annotations provided by the IRIM collaborative effort [5]. Because training SVM on large amount of data is expensive, we have

developed a specific training scheme to speed-up the process. This scheme is based on the use of Homogeneous Kernel Maps [34] which allow to approximate an SVM with a non-linear kernel by a linear SVM. For further speed-up, the feature values are quantized, with a non-linear quantization at 10,000 bins, so that the HKM can be precomputed only once. The order of the Homogeneous Kernel Maps is 5, so each scalar component is translated into a 11 dimension vector. To train the linear SVMs, we use a variation of the PEGASOS algorithm [30]. Our implementation is actually capable of training multiple models in parallel, corresponding to different concepts, or different values of the hyper-parameters. This allows to optimize the memory accesses to the stored feature vectors of the development set.

The TRECVID development data is split in four folds, we train on three folds and use the fourth one to select the best value of the hyper parameters. By rotating the folds, this leads to a set of four classifiers for each descriptor. We also apply Platt's normalization to transform the SVM score into a probability.

4.2 Early Fusion

We add to our set of descriptors some new descriptors obtained by early fusion of previous descriptor. First, we concatenate together the small size descriptors (color moments, wavelets, lbp, edge histograms) to form a single vector with 600 components. This is a simple concatenation without any selection.

We also build new descriptors by using a selection procedure over the components of all available descriptors. The criteria to select a component is the conditional entropy of the concepts given the value of the component. The components with the lowest entropy are selected and included in the resulting descriptor. The selection is done independently for each component. We know that this is not optimal, as once a component is selected, we should include its impact on the entropy of the remaining components, but the resulting computation would be too expensive. We consider different sizes for the number of components selected, 1,000, 4,096, and 8,192 components, so that we define three new descriptors in this early fusion scheme.

4.3 Late Fusion Schemes

Late Fusion is performed at various levels of the processing. For each descriptor, we have multiple models, depending on the folds used in the training, on the normalization into probabilities scores for a shot of the test data.

- we can average the scores coming from the models trained on different folds,
- we can average the probabilities coming from the models trained on different folds,
- we can compute a probability from the average score,
- we can combine the scores from different descriptors (using linear interpolation with an SVM),

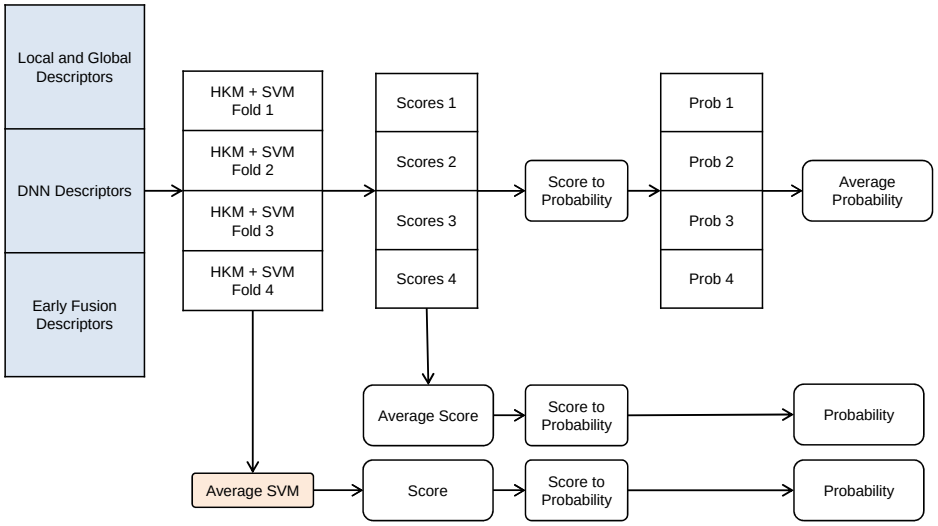

Figure 1: Late Fusion Schemes

- we can average the linear models before applying the probability conversion.

Those various schemes are illustrated in Figure 1.

4.4 Submitted runs

Our four runs are organized as follows:

1. **Run4**: For each descriptor, we average the unnormalized scores coming from models trained on various folds, then we compute a linear interpolation of the averages with a SVM. The linear interpolation is performed on a hierarchy of groups of descriptors, using four groups: all DNN descriptors, all DNN plus the four best non-DNN descriptors, all non-DNN descriptors, all descriptors. The scores are linearly interpolated within each group, then the four results are averaged for the final score.

2. **Run3**: Is it a similar process to Run4, but instead of averaging the scores, we first average the parameters of the SVM models. Then we also apply a probabilistic conversion. The resulting probabilities are again grouped in a hierarchy for linear interpolation.

3. **Run2**: It is again similar to Run4, but using the probabilities instead of the scores. The hierarchical interpolation process remains similar to the one in the previous runs.

4. **Run1**: It is a simple average of the scores obtained in the previous three runs. On our validation set, we found that it provided a slightly better performance.

5 SIN Results Analysis

Run	MAP 2014	MAP 2015
Run1	0.2175	0.2398
Run2	0.2025	0.2127
Run3	0.1315	0.2137
Run4	0.1151	0.2404

Figure 2: SIN Evaluation results for our runs in 2014 and 2015

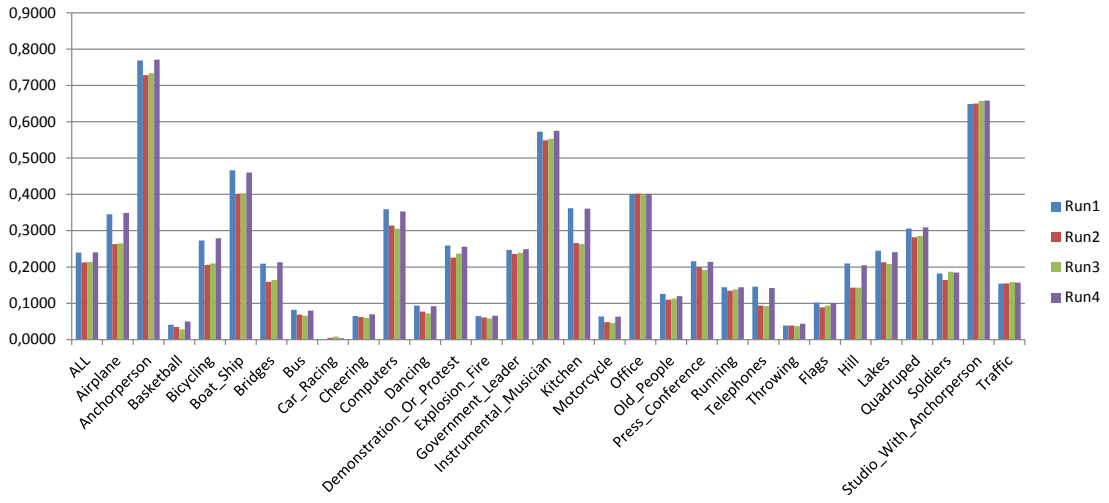

Figure 3: SIN Results on the test set evaluated by NIST

In Figure 2, we indicate the performances (MAP) of our runs for 2014 and 2015. Those figures are those provided by the manual evaluation performed by NIST. We can see that the best run

is Run4, meaning that the conclusions that we observed to select Run1 from the development data did not project to the 2015 test data. As most other groups in TRECVID, the use of other DNNs has produced an improvement of the performance. The comparison of the performance of our runs seems to indicate that the conversion to probabilities produced an overfit of the values, so that the probabilities on the test data are not as efficient as the scores. We plan to explore this issue further, for example by changing the validation scheme.

In figure 3 we display the comparative performance of the four runs on each of the concepts evaluated by NIST.

6 LNK framework motivation

The link between an anchor segment of a video and a target video segment in the collection can vary depending on the information the user is focusing on when choosing an anchor, and their general knowledge of the topic. As we do not have any information about potential users and their intentions when defining the anchors, our approach to extract these hyperlink targets automatically is based on all the available information, i.e. the audio-visual streams of the video collection. We use and connect both textual representations of the content with the results of automatic processing of the visual stream.

Recently several approaches did investigate potential to use the visual features when performing the task. In [29] the visual content was used to impose the segmentation units, while in [6] and [8] the visual concepts were used for reranking of the result list for the hyperlinking task. However, as the reliability of the extracted visual concepts and the types of the concepts themselves vary based on the training data and the task framework, it is still hard to transfer these systems output from one collection or task to another while keeping the same impact on improvement.

In our experiments we attempt to create this link between the textual content of an anchor and the visual features of the collection by incorporating the information about the words vectors distance into the confidence scores calculation. We take into account not only the transcript words corresponding to the anchor and words assigned to the visual concepts, but also their lexical context, calculated as close word vectors following the word2vec approach [18]. By expanding the list of terms for comparison by the lexical context, we attempt to deal with the potential mismatch of the terms used in the video and those describing the visual concepts, as the speakers in the videos might not directly describe the visual content, but it might be implied in the further lexical context of the topic of their speech.

We use the dataset of the Hyperlinking task at TRECVid 2015 [24] that contains both textual and visual descriptions of the required content, thus we can compare the influence of words vectors similarity for the cases when we establish the connection between the textual representation of the anchor and the visual content within the collection, and between the textual description of the visual request and the visual content within the collection.

7 LNK System Overview

To compare the impact of our approach, we create a baseline run that all further implementations are based upon.

First, we divide all the videos in the collection into segments of a fixed length of 120 seconds with a 30 seconds overlap step. We store the corresponding LIMSI transcripts [15] as the documents collection, and the information about the start of the first word after a pause longer than 0.5 seconds or a first switch of speakers as the potential jump-in point for each segment, as in [10].

Second, we use the open-source Terrier 4.0. Information Retrieval platform[1] [23] with a standard language modeling implementation [11], with default *lamda* value equal to 0.15, for indexing and retrieval. The resulting top 1000 segments for each of the 100 anchors represent the baseline result after the removal of the overlapping results.

Third, for these top 1000 segments we calculate a new confidence score that represents a combination of three values, see Equation 1: i) confidence score of the terms that are present both in the anchor ($C_{A\_w_i}$) and in the visual concepts extracted for the segment ($C_{VC\_w_i}$); ii) confidence score of the terms that are present both in the anchor ($C_{A\_w_i}$) and in lexical context of the visual concepts extracted for the segment ($C_{W2V4VC\_w_i}$); iii) confidence score of the terms that are present both in the lexical context of the anchor($C_{W2V4A\_w_i}$) and in the visual concepts extracted for the segment ($C_{VC\_w_i}$). We empirically chose to assign higher value (0.6) to the confidence score of the first type, as those are the words used in the transcripts and visual concepts, and lower equal values (0.2) for the scores using the lexical context, see Equation 1. We use the open-source implementation of the word2vec algorithm [2] with the pre-trained vectors trained on part of Google News dataset [3] (about 100 billion words), cf. [17]. We take the top 100 word2vec output for consideration, remove the stop words from both the query and the word2vec output, and run Porter Stemmer [25] on all lists for normalization.

Finally, the new confidence score values are used for the reranking of the initial results, these are filtered for the overlapping segments, and the jump-in points of the segments are used as start times.

$$
\begin{aligned}
ConfScore = & \frac{\sum_{i=1}^{N_{A\_VC}} (C_{A\_w_i} * C_{VC\_w_i})}{N_{A\_VC}} * 0.6 + \\
& + \frac{\sum_{i=1}^{N_{A\_W2V4VC}} (C_{A\_w_i} * C_{W2V4VC\_w_i})}{N_{A\_W2V4VC}} * 0.2 + \\
& + \frac{\sum_{i=1}^{N_{W2V4A\_VC}} (C_{W2V4A\_w_i} * C_{VC\_w_i})}{N_{W2V4A\_VC}} * 0.2
\end{aligned}
\tag{1}
$$

[1] http://www.terrier.org

[2] http://word2vec.googlecode.com/svn/trunk/

[3] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

Method ID	MAP			MAiSP
	overlap	bin	tolerance	
Baseline	0.2179	0.2130	0.1471	0.2020
Oxford	0.1154	0.1039	0.0699	0.1055
Leuven	0.1009	0.1075	0.1067	0.1067
CERTH	0.1248	0.1164	0.0725	0.1152

Table 1: LNK official metrics results

8 LNK Experimental Results

Table 7 show the official evaluation results of the created runs. The first line represent the results for the baseline run, while the other lines are naming the systems which produced the used visual concepts detection output: Oxford [7], Leuven [32] or CERTH [4].

9 Conclusions

This year EURECOM presented a set of systems for the Semantic INdexing and Video Hyperlinking Tasks. We introduced extra descriptors using Deep Neural Networks trained on ImageNet for the SIN task, and noticed that they produce an improvement on the performance of the detection of concepts in TRECVID shots. The effect of the various fusions schemes that we have used this year will deserve extra experiments to draw firm conclusions. LNK experiments showed that the connection of textual and visual features through the combination of confidence scores needs further analysis and tuning in order to improve over the baseline performance.

10 Acknowledgments

This work was supported by the European Commission's 7th Framework Programme (FP7) under FP7-ICT 287911 (LinkedTV); Bpifrance within the NexGen-TV Project, under grant number F1504054U.

References

[1] Color descriptors, http://koen.me/research/colordescriptors/.

[2] Local binary pattern, http://www.ee.oulu.fi/mvg/page/home.

[3] Vireo group in http://vireo.cs.cityu.edu.hk/links.html.

[4] E. E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Cervenková, D. Stein, S. Eickeler, J. L. R. García, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *Proceedings of the ACM International*

Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014, pages 1033–1036, 2014.

[5] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, mar 2008.

[6] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, 2013.

[7] K. Chatfield and A. Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *Computer Vision–ACCV 2012*, pages 432–446. Springer, 2013.

[8] S. Chen, M. Eskevich, G. J. F. Jones, and N. E. O'Connor. An investigation into feature effectiveness for multimedia hyperlinking. In *MultiMedia Modeling - 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part II*, pages 251–262, 2014.

[9] M. Eskevich and B. Huet. EURECOM @ SAVA2015: Visual Features for Multimedia Search. In *MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany, 2015.

[10] M. Eskevich and G. J. F. Jones. Time-based segmentation and use of jump-in points in DCU search runs at the search and hyperlinking task at mediaeval 2013. In *MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, 2013.

[11] D. Hiemstra. *Using language models for information retrieval.* PhD thesis, University of Twente, The Netherlands, 2001.

[12] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

[15] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer, 2008.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[18] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, May 2013.

[19] U. Niaz, B. Merialdo, and C. Tanase. EURECOM at TrecVid 2014: The semantic indexing task. In *TRECVID 2014, 18th International Workshop on Video Retrieval Evaluation, 10-12 November 2014, Orlando, USA*, Orlando, UNITED STATES, 11 2014.

[20] U. Niaz, M. Redi, C. Tanase, and B. Merialdo. EURECOM at TRECVID 2013: The light semantic indexing task. In *TRECVID 2013, 17th International Workshop on Video Retrieval Evaluation, 2013, National Institute of Standards and Technology, Gaithersburg, USA*, Gaithersburg, UNITED STATES, 11 2013.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971 –987, jul 2002.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[23] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[24] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quenot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[25] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[26] M. Redi and B. Mérialdo. Saliency moments for image categorization. In *ICMR'11, 1st ACM International Conference on Multimedia Retrieval, April 17-20, 2011, Trento, Italy*, 04 2011.

[27] M. Redi and B. Merialdo. Direct modeling of image keypoints distribution through copula-based image signatures. In *ICMR 2013, ACM International Conference on Multimedia Retrieval, April 16-19, Dallas, Texas, USA*, Dallas, ÉTATS-UNIS, 04 2013.

[28] B. Safadi, M. Sahuguet, and B. Huet. When Textual and Visual Information Join Forces for Multimedia Retrieval. In *Proceedings of International Conference on Multimedia Retrieval (ICMR'14)*, pages 265:265–265:272, Glasgow, United Kingdom, 2014.

[29] M. Sahuguet, B. Huet, B. Cervenková, E. E. Apostolidis, V. Mezaris, D. Stein, S. Eickeler, J. L. R. García, and L. Pikora. Linkedtv at mediaeval 2013 search and hyperlinking task. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.*, 2013.

[30] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[32] T. Tommasi, T. Tuytelaars, and B. Caputo. A testbed for cross-dataset analysis. *CoRR*, abs/1402.5923, 2014.

[33] 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*, 32(9):1582–1596, 2010.

[34] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480 – 492, 2012.