# Linear SVM for new Pyramidal Multi-Level Visual only Concept Detection in CLEF 2010 Challenge

Sébastien PARIS[1], Hervé GLOTIN[2]
[1]*LSIS DYNI, Univ Paul Cézanne, av Escadrille Normandie-Niemen,13397 MARSEILLE CEDEX 20*
[2]*LSIS DYNI, Univ du Sud-Toulon-Var, av de l'Université - BP20132, 83957 LA GARDE*
*sebastien.paris@lsis.org, glotin@univ-tln.fr*

**Abstract.** For the Visual Concept Detection of CLEF 2010 Challenge, using only visual information, we propose a novel multi-level spatial pyramidal (sp) features : the spELBP (Extended Local Binary Pattern). In this paper we first present these features and few others that are similar : the spELBOP (Extended Local Binary Orientation Pattern), and the spHOEE (Histogram of Oriented Edge Energy). Then we discuss why our features feed state-of-the-art linear SVM algorithms for the Detection of Concept. Our scores are ranked, over the 15 participating teams, $8^{th}$ according to the F-mesure evaluation, and $9^{th}$ according to the MAP evaluation. We compare each topic score to the best system, and we finaly discuss on further extension of our approach [1].

## 1 Introduction

The VCDT 2010 challenge (see [NH10]) consists in detection of visual concept.

Basically, this challenge can be considered as a supervised classification problem, more precisely by training models on efficient features with a "one-against-all" approach. In recent years in computer vision, in order to reduce the semantic gap in object categorization problem, two popular approaches have emerged offering efficient performances. The first one, *a.k.a.* "Bag of Words" (BoW) (see [ZZY$^+$,YYGH]), consists in building a dictionary of visual words given a large pool of feature vectors, usually some SIFT descriptors [Low]. SIFT descriptors can be computed over a regular spatial grid or on interest point outputs of specific detectors (corners, edges, blobs, ...) such Harris or Lowe detectors [HS88,Low]. Following a dictionary learning step usually done by a vector quantification of all the total amount of feature vectors. The vector quantification is usually done by a K-means or GMM algorithms [YYGH]. More efficient dictionaries can be retrieved with sparse learning tools [WYKY$^+$].

The second approach is based on Local Binary Pattern (LBP) descriptors (*a.k.a.* CENTRIST in [WR]). The feature vector is defined by occurrences of each 256 patterns encoding the neighborhood relation to a central pixel.

[1] This work has been supported by ANR-06-MDCA-002 AVEIR ANR

For both approaches, adding a multi-level pyramidal architecture, permits to improve considerably the performances. This technic divides the image in sub-windows and weights adequately each corresponding feature vectors before concatenation (see [LSP]). The price of this kind of architecture is to deal with much larger vectors as input of classifiers. Large-scales binary supervised classification problems arise naturally with theses descriptors.

The next section describes more precisely the descriptors we developped in the challenge, especially the novel descriptor spELBOP. The fourth section overviews the large-scale binary classifier we use : the linear SVM Classifier TRON (L2 regularized with a L2 loss function).

## 2 Pyramidal Multi-Level Features

For each of the three following descriptors, a spatial pyramid architecture is used to divide the entire image $\boldsymbol{I}$ into $Ns$ possibly overlapping sub-windows. More precisely, a $L$ levels pyramid is defined for $l = 1, \ldots, L$, where image $\boldsymbol{I}$ of size $Ny \times Ny$ is divided into possibly overlapping sub-windows of size $h_l \times w_l$. Histograms are computed for each sub-windows and weighted by $c_l = \left(\frac{\max_{j=1,\ldots,L}\{h_j\}}{h_l}\right).\left(\frac{\max_{j=1,\ldots,L}\{w_j\}}{w_l}\right)$. Finally, concatenation of the $Ns$ weighted histograms defines the global feature vector. In our implementation, $h_l = \lfloor Ny.r_{y,l} \rfloor$ and $w_l = \lfloor Nx.r_{x,l} \rfloor$ where $r_{y,l}$ and $r_{x,l}$ are elements of vectors $\boldsymbol{r}_y$ and $\boldsymbol{r}_x$. Shifts in x-y axis are defined by integers $\delta_{y,l} = \lfloor Ny.d_{y,l} \rfloor$ and $\delta_{x,l} = \lfloor Nx.d_{x,l} \rfloor$ where $d_{y,l}$ and $d_{x,l}$ are elements of vectors $\boldsymbol{d}_y$ and $\boldsymbol{d}_x$ respectively. Overlapping windows can be obtained if $d_{y,l} \leq r_{y,l}$ and/or $d_{x,l} \leq r_{x,l}$. The total number of sub-windows is equal to $Ns = \sum_{l=1,\ldots,L} \lfloor \frac{(1-r_{y,l})}{(d_{y,l}+1)} \rfloor . \lfloor \frac{(1-r_{x,l})}{(d_{x,l}+1)} \rfloor$.

### 2.1 The spHOEE Feature

Following [DT,MBM], a histogram of the L1-normalized orientation edge energy filter responses is constructed for the $No$ different orientations. Theses responses are obtained by convolution of the gray image with two odd elongated oriented filters (horizontal and vertical gradients) at scale $\sigma$. L1-normalized magnitudes with a block of size $h_n \times w_n$ and signed angles are computed from theses gradients. Each $Ns$ sub-window histogram is computed efficiently thanks to the integral histogram method. The total dimension of the feature vector is $d = Ns.No$. The spHOEE feature (*a.k.a.* spHOG in [MBM,MB]) offers state-of-the-art performances in databases such *CALTECH 101* or *INRIA pedestrians*.

### 2.2 The novel spELBP Feature

Local Binary Pattern (LBP) are powerful parametric descriptors encoding relation between intensity of a central pixel and intensities of its 8 adjacent neighbors (see [LZL+07]). Widely used in face recognition (see [SGM09]), LBP shows also

their efficiency in scene categorization ([WR]) compared to BoW with SIFT descriptors. In [LZL+07], a multi-scale extension (MSLBP) consists in encoding relation of a central block of pixels of size $s \times s$ with its 8 neighbors capturing more global details. Each block area is computed with the help of the integral image. We propose here a spatial pyramid architecture for the MSLBP so-called spELBP. This novel descriptor captures details of the scale $s$ at given sub-windows location. Let $S$ the number of scales, the total dimension of the spELBP descriptor is $d = 256.Ns.S$.

### 2.3 The spELBOP Feature

This novel descriptor is derived from the two last. Here instead of encoding the raw pixel values of a block of size $s$, we propose to encode the orientations of the block. As with the spHOEE features, orientations are retrieved by i) applying convolution with the two odd elongated oriented filters at scale $\sigma$ and ii) computing the signed angles. The total dimension of the spELBOP descriptor is the same as the latter, *i.e.* $d = 256.Ns.S$.

## 3 Large Scale SVM

Learning a topic (a room) with the one-against-all approach is equivalent to a binary supervised classification task. We deal with a training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}, i = 1, \ldots, N$ where $\boldsymbol{x}_i \subset \mathcal{R}^d$ represents a feature vector and $y_i \in \{-1, 1\}$ its corresponding label. Max-margin classifiers like SVM are known to offer state-of-the-art performances. However with high dimension feature vectors and numerous examples, training SVM can be too computational expensive ($\sim O(dN^3)$). For large scale problems, one alternative is to use a max-margin linear classifier which offers often the same amount of performances than the non-linear version [LWK,WYKY+]. The linear SVM used here consists in finding the hyperplane parameter $\boldsymbol{w}$ minimizing the sum of a L2 loss function and a L2 regulation term:

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N} \max\left(1 - y_i\boldsymbol{w}^T\boldsymbol{x}_i, 0\right)^2 \right\}. \quad (1)$$

In [LWK], the problem is solved with a Trust Region Newton algorithm (TRON). We use the modified version of TRON proposed by ([MBM]) managing dense features.

## 4 Results at VCDT 2010

In the CLEF VCDT 2010 task, preliminary tests on the development set resulted in selecting the spELBP feature for our final system. We choose a $L = 3$ levels pyramid $\boldsymbol{r}_x = \boldsymbol{r}_y = \boldsymbol{d}_x = \boldsymbol{d}_y = \left[1, \frac{1}{2}, \frac{1}{4}\right]^T$ leading to $Ns = 42$ sub-windows and a total of $d = 10752$ dimensions for this feature.

For each topic, hyperparameters $C$ or $\lambda$ of classifiers are tuned with a 5 cross-validation by minimizing the Balanced Error Rate (BER). Then models are learned on entire training sets.

All our run are visual only runs (we only use the image pixels to detect the topics).

Over the 15 participating teams, and for the example-based evaluation applying the F-Measure, our system has the 8th rank. For the seconde evaluation, our system has the 9th rank according to the concept-based Mean Average Precision (MAP) (this measure showed better characteristics than the EER and AUC in a recent study).

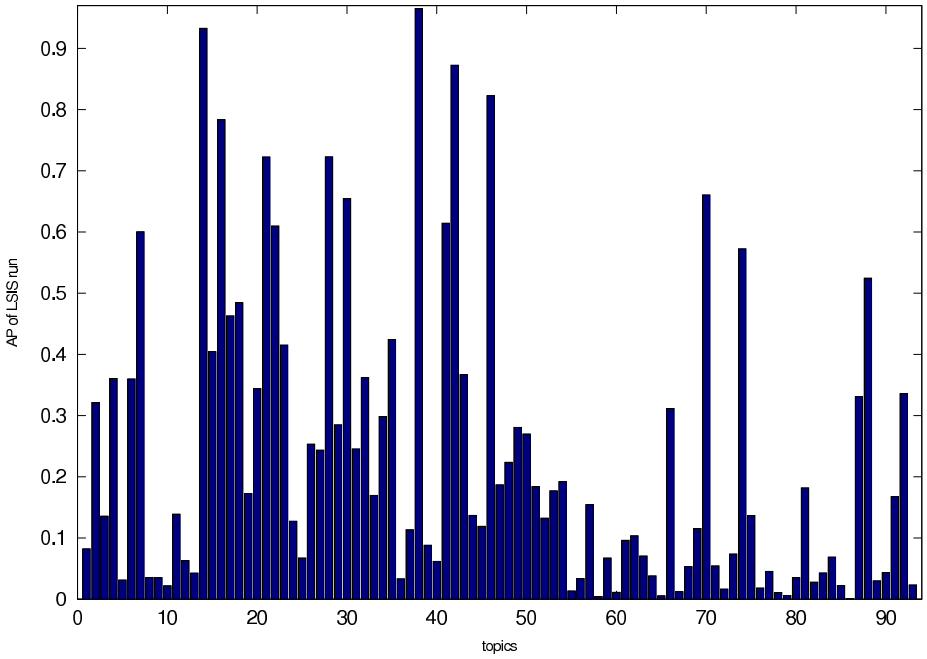


**Fig. 1.** Average Precision of the best LSIS run for all the topics.

Score details are given at :
http://www.imageclef.org/2010/PhotoAnnotationExampleEvaluationResults
We give below the team best run sorted list with the rank, Fmesure, Team and Run ID :
1 0.680070 ISIS_1276866836402_uva-isis-mkl-mixed-mixed.txt_binary.txt
2 0.639441 XRCE_1277144880578_xrce_SVM_EF_Visual.txt_binary.txt
3 0.634064 HHI_1277376108118_ic_10_test_s_eiq_space.txt_binary.txt
4 0.596141 IJS_1277145320629_final_ijs_feit_run1.txt_binary.txt
5 0.581652 LEAR_1277147818075_lear_TP_Visual_CVF_0_D6.txt_binary.txt

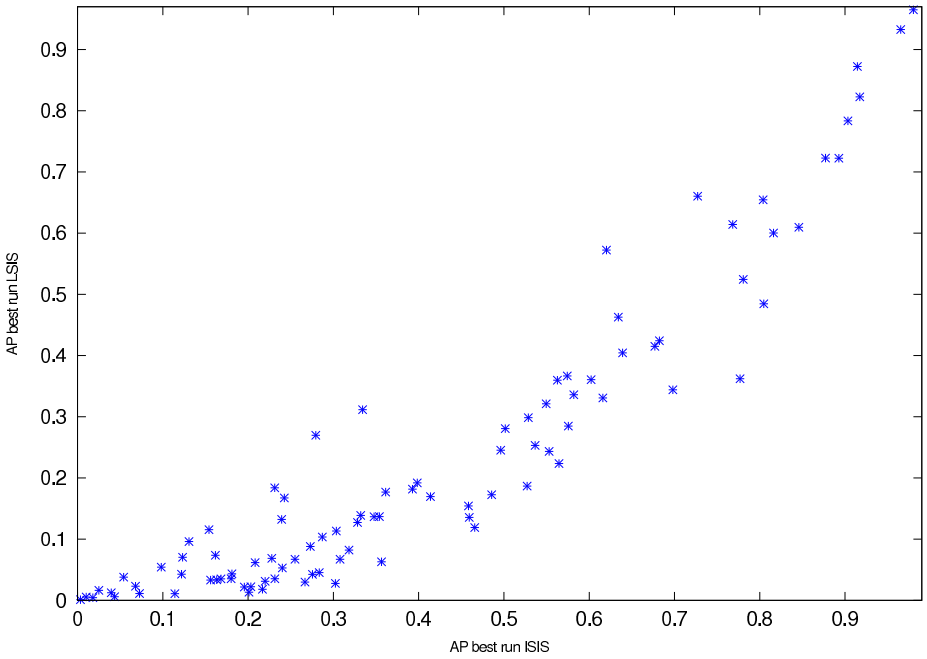


**Fig. 2.** Average Precision of the best ISIS run versus best LSIS run for all the topics.

6 0.558674 MEIJI_1276044391982_AUTO_VisualOnly_BagOfVisualWords_of_meiji.txt_binary.txt
7 0.530841 Romania_1276777329876_run_CR+lapl2inv.txt_binary.txt
8 0.530317 LSIS_1277226430977_DYNI_LSIS_RUN2_COPIE.txt_binary.txt
9 0.482390 WROCLAW_1277754391699_imageClef2010-grid20x20-xy_rgb_dev_hes.quick_matrix...
10 0.476983 LIG_1277153756343_clefResults.txt_binary.txt
11 0.450934 CEALIST_1277046611397_cealist_fastSB_6600.submission.txt_binary.txt
12 0.427661 BPACAD_1277129525816_bp_acad_hoggmm.txt_binary.txt
13 0.224987 MLKD_1277149221968_Visual2.txt_binary.txt
14 0.208564 INSUNHIT_1277043267771_finalresult50.txt_binary.txt
15 0.174392 UPMC_1277139769194_output_multiviewFINAL.txt_binary.txt

The best MAP run is also ISIS run with 0.407, while the best LSIS run MAP equals 0.234.

The figure 1 gives the Average Precision (AP) for each topic for the best LSIS run. In order to compare our scores to the state of the art, we plot in figure 2 the AP of the best run of the challenge versus our. We then clearly see that the scores are well correlated, and that LSIS AP are below (or for some topics equal) to the ISIS AP. This loss of precision for the LSIS run is figured for each topic in 3. We see then that the difference is about 0.15 points of AP in average.

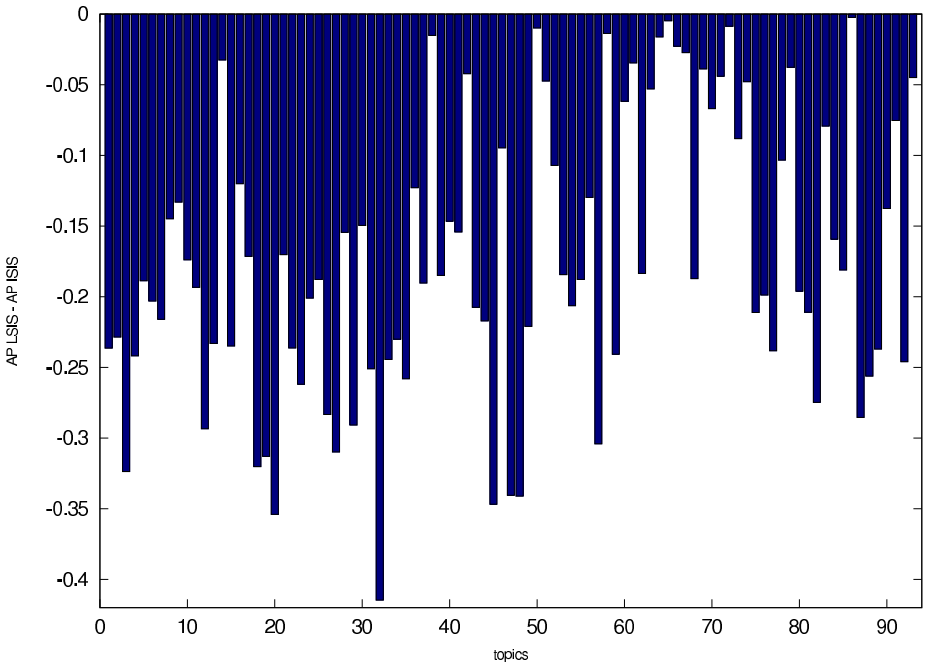


**Fig. 3.** Difference, for each topic, of the Average Precision between the best LSIS run and the best run of VCDT2010 (from ISIS).

## 5 Conclusion and perspectives

In this paper, the VCDT 2010 challenge is solved by a large-scale classifier trained on multi-level descriptors and is ranked 8th for the F-mesure evaluation over the 15 teams. Training theses classifiers is extremely fast compared to the classical SVM counterpart. The Average Precision of our system is below the best run of the challenge with a nearly constant decrease of around 0.15. We could then assume that our approach could be globaly improved, as it seems to generalize well to all topics. Further improvements can be expected by adding denseSIFT descriptor with sparse learning and spatial pooling (see [WYKY$^+$]), and using Multiple Kernel Learning method for fusionning features.

## References

[DT] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc 4th Alvey Vision Conf*, 1988.

[Low] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV'99*.

[LSP] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'06*.

[LWK] Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region newton method for logistic regression. *J. Mach. Learn. Res.*, 9.

[LZL+07] ShengCai Liao, XiangXin Zhu, Zhen Lei, Lun Zhang, and Stan Z. Li. Learning multi-scale block local binary patterns for face recognition. In *ICB*, 2007.

[MB] S. Maji and A.C. Berg. Max-margin additive classifiers for detection. *ICCV'09*.

[MBM] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *CVPR'08*, June.

[NH10] S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *Working notes of CLEF 2010*, 2010.

[SGM09] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27(6), 2009.

[WR] Jianixn Wu and James M. Rehg. Where am i: Place instance and category recognition using spatial pact. *CVPR'08*.

[WYKY+] Jinjun Wang, Jianchao Yang, Fengjun Lv Kai Yu, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. *CVPR'10*.

[YYGH] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR'09*.

[ZZY+] Xi Zhou, Xiaodan Zhuang, Shuicheng Yan, Shih-Fu Chang, Mark Hasegawa-Johnson, and Thomas S. Huang. Sift-bag kernel for video event analysis. In *MM'08*. ACM.