Pyramidal Multi-Level Features for the RobotVision@ICPR 2010 Challenge

Sébastien PARIS\textsuperscript{1}, Hervé GLOTIN\textsuperscript{2}

\textsuperscript{1}LSIS DYNI, Univ Paul Cézanne, av Escadrille Normandie-Niemen,13397 MARSEILLE CEDEX 20
\textsuperscript{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

This paper combines and proposes two novel multi-level spatial pyramidal (sp) features: spELBP (Extended Local Binary Pattern), spELBOP (Extended Local Binary Orientation Pattern) and spHOEE (Histogram of Oriented Edge Energy). These features feed state-of-the-art SVM algorithms for the localization of a robot in indoor environments. Two tasks are associated with the RobotVision@ICPR 2010 Challenge, the first one uses only a frame of stereoscopic images, the second takes into account the dynamics of the robot for improving results. Our scores are ranked 3\textsuperscript{rd} for Task1 and 1\textsuperscript{st} for Task2\textsuperscript{1}.

1. Introduction

The RobotVision@ICPR 2010 challenge (see [1]) consists in building a system permitting the localization of a binocular robot manoeuvring in indoor environments. To train robot's models, two databases of stereoscopic images named "easy" and "hard" are provided and labeled into 9 topics associated with the visited rooms during robot's path: Elevator(1), Corridor(2), Kitchen(3), LargeOffice1(4), LargeOffice2(5), SmallOffice2(6), StudentOffice(7), Lab(8), PrinterArea(9) (see Fig.1 for a example). Theses two databases differ from the number of training images, illumination conditions and robot's paths taken. Two others databases are also provided, \textit{i.e.} the labeled "validation" database containing also 9 topics permitting evaluation of the trained system and the unlabeled "test" database containing images of 13 visited rooms. For this latter, the 4 extra rooms must be classified as a 10\textsuperscript{th} class "Unknown". Each participant must return two predicted labels sequences of the "test" database where models had been trained on the "easy" and "hard" databases.

Basically, this challenge can be considered as a supervised classification problem, more precisely by training 9 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, \textit{a.k.a.} "Bag of Words" (BoW) (see [16, 15]), consists in building a dictionary of visual words given a large pool of feature vectors, usually some SIFT descriptors [7]. 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 [3, 7]. 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 [15]. More efficient dictionaries can be retrieved with sparse learning tools [13].

The second approach is based on Local Binary Pattern (LBP) descriptors (a.k.a. CENTRIST in [14]). The feature vector is defined by occurrences of each 256 patterns encoding the neighborhood relation to a central pixel.

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 [4]). 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 three descriptors used in the challenge, especially the novel descriptor spELBP. The fourth section overviews two large-scale binary classifiers used for task 1: i) the linear SVM Classifier TRON (L2 regularized with a L2 loss function) and ii) the approximate additive kernel SVM PIWSGD (L2 regularized with a L1 loss function). The following section presents the Hidden Markov Models (HMM) used in task 2 in order to take into account the dynamic of the robot. Results of these two tasks are presented in section 5.

2. Pyramidal Multi-Level Features

For each of the three following descriptors, a spatial pyramid architecture is used to divide the entire image \( I \) into \( N_s \) possibly overlapping sub-windows. More precisely, a \( L \) levels pyramid is defined for \( l = 1, \ldots, L \), where image \( I \) of size \( N_y \times N_y \) is divided into possible overlapping sub-windows of size \( h_l \times w_l \). Histograms are computed for each sub-windows and weighted by \( c_l = \left( \frac{\max(h_l)}{h_l} \right) \left( \frac{\max(w_l)}{w_l} \right) \). Finally, concatenation of the \( N_s \) weighted histograms defines the global feature vector. In our implementation, \( h_l = \lfloor N_y r_y l \rfloor \) and \( w_l = \lfloor N_x r_x l \rfloor \) where \( r_y l \) and \( r_x l \) are elements of vectors \( r_y \) and \( r_x \). Shifts in \( x-y \) axis are defined by integers \( \delta_{y,l} = \lfloor N_y d_{y,l} \rfloor \) and \( \delta_{x,l} = \lfloor N_x d_{x,l} \rfloor \) where \( d_{y,l} \) and \( d_{x,l} \) are elements of vectors \( d_y \) and \( d_x \) 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 \( N_s = \sum_{l=1}^{L} \frac{1}{(d_{y,l} + 1)} \frac{1}{(d_{x,l} + 1)} \).

2.1. The spHOEE Feature

Following [2, 9], a histogram of the L1-normalized orientation edge energy filter responses is constructed for the \( N_o \) different orientations. These 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 these gradients. Each \( N_s \) sub-window histogram is computed efficiently thanks to the integral histogram method. The total dimension of the feature vector is \( d = N_s N_o \). The spHOEE feature (a.k.a. spHOG in [9, 8]) offers state-of-the-art performances in databases such CALTECH 101 or INRIA pedestrians.

2.2. The 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 [5]). Widely used in face recognition (see [12]), LBP shows also their efficiency in scene categorization ([14]) compared to BoW with SIFT descriptors. In [5], a multi-scale extension (MSLBP) consists in encoding relation of a central block of pixels of size \( s \times s \) with its 8 neighbors (see Fig. 3) 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 \( d = 256 N_s S \).

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. N. S. S.

3. Large Scale SVM for Task 1

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 $\mathcal{D} = \{x_i, y_i\}, i = 1, \ldots, N$ where $x_i \in \mathbb{R}^d$ 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 [6, 13]. The linear SVM used here consists in finding the hyperplane parameter $w$ minimizing the sum of a L2 loss function and a L2 regulation term:

$$
\min_{w} \left\{ \frac{1}{N} \sum_{i=1}^{N} \max(1 - y_i w^T x_i, 0)^2 \right\}.
$$

(1)

In [6], the problem is solved with a Trust Region Newton algorithm (TRON). We use the modified version of TRON proposed by ([9]) managing dense features.

Another alternative is to use additive kernel (like histogram intersection kernel) where the evaluation function can be written as $f(x) = \sum_{i=1}^{d} f_i(x_i)$. In [8], they proposed to use special data encoding permitting to approximate linearly $f^{W'}(\hat{x}) \approx \hat{w}^T \hat{x}$. After encoding, the problem is equivalent to find the hyperplane parameter $\hat{w}$ minimizing the sum of a L1 loss function and a L2 regulation term:

$$
\min_{\hat{w}} \left\{ \lambda \hat{x}^T H \hat{x} + \frac{1}{N} \sum_{i=1}^{N} \max(1 - y_i \hat{w}^T \hat{x}_i, 0) \right\},
$$

(2)

where $H$ is a tridiagonal matrix. This minimization is performed with an on-line stochastic sub-gradient descend algorithm so-called PWLSGD for piece-wise linear stochastic gradient descent.

For both classifiers, the a posteriori probability $\Pr(y = i|x)$ can be retrieved after fitting sigmoid parameters on classifier’s outputs $(f(x_i)), i = 1, \ldots, N$ trained for each topic. For the "Unknown" topic ($y = 10$), we chose $\Pr(y = 10|x) = \eta$ where $\eta$ is optimized on the validation set.

4. A Hidden Markov Model for Task 2

In order to smooth the raw labels sequence $\{y_k\}, k = 1, \ldots, K$ ($K$ denotes the number of frames) outputs of previous classifiers, we propose to use a HMM, more precisely the Forward-Backward algorithm (see [11, 10]). The state transition probabilities matrix $A$ ($10 \times 10$) is designed according to the floorplan where each room is connected to the corridor (with label $y = 2$) and other rooms are not directly connected each others. $A = \{a_{ij}\}$ is defined by:

$$
a_{ij} = \Pr(y_k = i|y_{k-1} = j) = \begin{cases} 
1 - \lambda, & i = j \\
\lambda, & i = 2, j \neq i \\
\frac{\lambda}{4}, & i \neq j, j = 2 \\
0, & \text{else}
\end{cases}
$$

(3)

with $\lambda = \frac{1}{r}$, where $r$ represents the mean sejour in current state. The conditional measurements probabilities $\Pr(x_k|y_k = i)$ are proportional to probabilist outputs of SVM trained on each topic during task 1.

5. Results for RobotVision@ICPR 2010

For spHOEE feature, we choose a $L = 4$ levels pyramid $r_x = r_y = \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, $d_x = d_y = \frac{1}{16}$, $\alpha = 2$, $No = 12$, leading to $Ns = 540$ sub-windows, and a total of $d = 6480$ dimensions. For spELBP and spELBOP features, we choose a $L = 3$ levels pyramid $r_x = r_y = d_x = d_y = \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ leading to $Ns = 42$ sub-windows and a total of $d = 10752$ dimensions for these features.

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 "Easy" and "Hard".

We present in Tab. 1 the performances evaluated for the "test" sequence ($K = 2551$) in terms of Error Rate (ER), score criterion used for RobotVision@ICPR 2010 ($\frac{K}{d} \leq SRV \leq K$, see [1]) and mean area under curve ($AUC$ for the 9 first topics), for each descriptor and for the two used classifiers. We see that for the "Easy" database, whatever the descriptors, PLWSGD improves the results. The spELBP or spELBOP descriptors outperform spHOEE.

In Tab. 2, we present results of the fused system where fusion rules (mean of selected a posteriori probabilities and features) had been learned on the "validation" set for the four following configurations: "Easy-Task1", "Easy-Task2", "Hard-Task1" and
Table 1. TRON/PLWSGD results (Task 1)

<table>
<thead>
<tr>
<th>Set</th>
<th>Classifier</th>
<th>ER</th>
<th>SRV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>TRON</td>
<td>0.328</td>
<td>1292.5</td>
<td>0.9251</td>
</tr>
<tr>
<td>Easy</td>
<td>PWLSGD</td>
<td>0.216</td>
<td>1724.5</td>
<td>0.9687</td>
</tr>
<tr>
<td>Hard</td>
<td>TRON</td>
<td>0.398</td>
<td>1025.5</td>
<td>0.8471</td>
</tr>
<tr>
<td>Hard</td>
<td>PWLSGD</td>
<td>0.371</td>
<td>1129.0</td>
<td>0.8394</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>Classifier</th>
<th>ER</th>
<th>SRV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>TRON</td>
<td>0.224</td>
<td>1691.5</td>
<td>0.9683</td>
</tr>
<tr>
<td>Easy</td>
<td>PWLSGD</td>
<td>0.156</td>
<td>1952.5</td>
<td>0.9783</td>
</tr>
<tr>
<td>Hard</td>
<td>TRON</td>
<td>0.338</td>
<td>1255.0</td>
<td>0.8855</td>
</tr>
<tr>
<td>Hard</td>
<td>PWLSGD</td>
<td>0.350</td>
<td>1211.5</td>
<td>0.8452</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>Classifier</th>
<th>ER</th>
<th>SRV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>TRON</td>
<td>0.219</td>
<td>1711.0</td>
<td>0.9699</td>
</tr>
<tr>
<td>Easy</td>
<td>PWLSGD</td>
<td>0.215</td>
<td>1727.5</td>
<td>0.9621</td>
</tr>
<tr>
<td>Hard</td>
<td>TRON</td>
<td>0.388</td>
<td>1066.0</td>
<td>0.8416</td>
</tr>
<tr>
<td>Hard</td>
<td>PWLSGD</td>
<td>0.393</td>
<td>1043.5</td>
<td>0.8395</td>
</tr>
</tbody>
</table>

Table 2. Fusion results (Tasks 1 and 2)

<table>
<thead>
<tr>
<th></th>
<th>ER</th>
<th>Easy</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Fusion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>0.1477</td>
<td>0.2994</td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>0.0619</td>
<td>0.2567</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Easy</th>
<th>Hard</th>
<th>Easy+Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>1985.5</td>
<td>1405.0</td>
<td>3390.5</td>
</tr>
<tr>
<td>Task 2</td>
<td>2314.0</td>
<td>1568.5</td>
<td>3882.5</td>
</tr>
</tbody>
</table>

6. Conclusion and perspectives

In this paper, the RobotVision@ICPR 2010 challenge is solved by fusion of two large-scale max-margin classifier's predictions trained on multi-level descriptors and is ranked third for the Task 1. Training these classifiers is extremely fast compared to the classical SVM counterpart. spELBP and spELBOP descriptors improve performances up to 28% in ER compared to spHOEE. For Task 2, smoothing the prediction sequence by a Forward-Backward algorithm improves robustness of the system. Further improvements can be expected by adding denseSIFT descriptor with sparse learning and spatial pooling (see [13]), and using Multiple Kernel Learning method for fusingion features.

References