Analysis of the Multidimensional Scale Saliency Algorithm and its Application to Texture Classification

Oral Talk
Pablo Suau & Francisco Escolano (UA) Spain

S+S SPR
August 18 - 20 2010
Cesme, TURKEY
Interest Points

Background. From the classical Harris detector, a big bang of interest-point detectors ensuring some sort of invariance to zoom/scale, rotation and perspective distortion has emerged since the proposal of SIFT detector and descriptors [Lowe, 04]. These detectors, typically including a multi-scale analysis of the image, include Harris Affine, MSER [Matas et al, 02] and SURF [Bay et al, 08]. The need of an intensive comparison of different detectors (and descriptors) mainly in terms of spatio-temporal stability (repeatability, distinctiveness and robustness) is yet a classic challenge [Mikolajczyk et al, 05]. Stability experiments are key to predict the future behavior of the detector/descriptor in subsequent tasks (bag-of-words recognition, matching,...)
Local Saliency, in contrast to global one [Ullman, 96], means local distinctiveness (outstanding/popout pixel distributions over the image) [Julesz, 81][Nothdurft, 00]. In Computer Vision, a mild IT definition of local saliency is linked to visual unpredictability [Kadir and Brady, 01]. Then a salient region is locally unpredictable (measured by entropy) and this is consistent with a peak of entropy in scale-space. Scale-space analysis is key because we do not know the scale of regions beforehand. In addition, isotropic detections may be extended to affine detectors with an extra computational cost. (see Alg. 1).
**Alg. 1:** Kadir and Brady scale saliency algorithm

**Input:** Input image I, initial scale $s_{min}$, final scale $s_{max}$

for each pixel $x$

for each scale $s$ between $s_{min}$ and $s_{max}$ do

calculate local entropy $H_D(s, x) = -\sum_{i=1}^{L} P_{s,x}(d_i) \log_2 P_{s,x}(d_i)$

end

choose the set of scales at which entropy is a local maximum

$S_p = \{ s : H_D(s - 1, x) < H_D(s, x) > H_D(s + 1, x) \}$

for each scale $s$ between $s_{min}$ and $s_{max}$ do

if $s \in S_p$ then

entropy weight calculation by means of a self-dissimilarity measure in scale space

$W_D(s, x) = \frac{s^2}{2s-1} \sum_{i=1}^{L} | P_{s,x}(d_i) - P_{s-1,x}(d_i) |$

entropy weighting $Y_D(s, x) = H_D(s, x) W_D(s, x)$

end

end

end

**Output:** A disperse three dimensional matrix containing weighted local entropies for all pixels at those scales where entropy is peaked
Entropy Saliency Detector (and 3)

Figure: Isotropic (top) vs Affine (bottom) detections.
Scale-space analysis, is, thus, one of the bottlenecks of the process. However, having a prior knowledge of the statistics of the images being analyzed it is possible to discard a significant number of pixels, and thus, avoid scale-space analysis.

**Working hypothesis**

If the local distribution around a pixel at a scale $s_{max}$ is highly homogeneous (low entropy) one may assume that for scales $s < s_{max}$ it will happen the same. Thus, scale-space peaks will not exist in this range of scales. [Suau and Escolano, 08].

Inspired in statistical detection of edges [Konishi et al., 03] and contours [Cazorla & Escolano, 03] and also in contour grouping [Cazorla et al., 02].
Relative entropy and threshold, a basic procedure consists of computing the ratio between the entropy at $s_{\text{max}}$ and the maximum of entropies for all pixels at their $s_{\text{max}}$.

Filtering by homogeneity along scale-space

1. Calculate the local entropy $H_D$ for each pixel at scale $s_{\text{max}}$.
2. Select an entropy threshold $\sigma \in [0, 1]$.
3. $X = \{x \mid \frac{H_D(x, s_{\text{max}})}{\max_x \{H_D(x, s_{\text{max}})\}} > \sigma\}$
4. Apply scale saliency algorithm only to those pixels $x \in X$.

What is the optimal threshold $\sigma$?
Images belonging to the same image category or environment share similar intensity and texture distributions, so it seems reasonable to think that the entropy values of their most salient regions will lay in the same range.

**On/Off distributions**

The $p_{on}(\theta)$ defines the probability of a region to be part of the most salient regions of the image given that its relative entropy value is $\theta$, while $p_{off}(\theta)$ defines the probability of a region to do not be part of the most salient regions of the image. Then, the maximum relative entropy $\sigma$ being $p_{on}(\sigma) > 0$ may be choosen as an entropy threshold for that image category by finding a trade-off between false positives and negatives.
Learning and Chernoff Information (4)

Figure: On(blue)/Off(red) distributions and thresholds.
Chernoff Information

The expected error rate of a likelihood test based on \( p_{\text{on}}(\phi) \) and \( p_{\text{off}}(\phi) \) decreases exponentially with respect to \( C(p_{\text{on}}(\phi), p_{\text{off}}(\phi)) \), where:

\[
C(p, q) = - \min_{0 \leq \lambda \leq 1} \log \left( \sum_{j=1}^{J} p^{\lambda}(y_j) q^{1-\lambda}(y_j) \right)
\]

A related measure is Bhattacharyya Distance (Chernoff with \( \lambda = 1/2 \)):

\[
BC(p, q) = - \log \left( \sum_{j=1}^{J} p^{1/2}(y_j) q^{1/2}(y_j) \right)
\]
A threshold $T$ must be chosen for an image class so any pixel from an image belonging to the same image class may be discarded if $\log(p_{on}(\theta)/p_{off}(\theta)) < T$. Being $T$

$$-D(p_{off}(\theta)||p_{on}(\theta)) < T < D(p_{on}(\theta)||p_{off}(\theta))$$

and $D(.||.)$ the Kullback-Leibler divergence:

$$D(p||q) = \sum_{j=1}^{J} p(y_j) \log \frac{p(y_j)}{q(y_j)}$$
Learning and Chernoff Information (7)

Figure: Filtering for different image categories, $T = 0$
<table>
<thead>
<tr>
<th>Test set</th>
<th>Chernoff</th>
<th>T</th>
<th>% Points</th>
<th>% Time</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplanes_side</td>
<td>0.415</td>
<td>-4.98</td>
<td>30.79%</td>
<td>42.12%</td>
<td>0.0943</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>60.11%</td>
<td>72.61%</td>
<td>2.9271</td>
</tr>
<tr>
<td>background</td>
<td>0.208</td>
<td>-2.33</td>
<td>15.89%</td>
<td>24.00%</td>
<td>0.6438</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>43.91%</td>
<td>54.39%</td>
<td>5.0290</td>
</tr>
<tr>
<td>bottles</td>
<td>0.184</td>
<td>-2.80</td>
<td>9.50%</td>
<td>20.50%</td>
<td>0.4447</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>23.56%</td>
<td>35.47%</td>
<td>1.9482</td>
</tr>
<tr>
<td>camel</td>
<td>0.138</td>
<td>-2.06</td>
<td>10.06%</td>
<td>20.94%</td>
<td>0.2556</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>40.10%</td>
<td>52.43%</td>
<td>4.2110</td>
</tr>
<tr>
<td>cars_brad</td>
<td>0.236</td>
<td>-2.63</td>
<td>24.84%</td>
<td>36.57%</td>
<td>0.4293</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>48.26%</td>
<td>61.14%</td>
<td>3.4547</td>
</tr>
<tr>
<td>cars_brad_bg</td>
<td>0.327</td>
<td>-3.24</td>
<td>22.90%</td>
<td>34.06%</td>
<td>0.2091</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>57.18%</td>
<td>70.02%</td>
<td>4.1999</td>
</tr>
<tr>
<td>faces</td>
<td>0.278</td>
<td>-3.37</td>
<td>25.31%</td>
<td>37.21%</td>
<td>0.9057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>54.76%</td>
<td>67.92%</td>
<td>8.3791</td>
</tr>
<tr>
<td>google_things</td>
<td>0.160</td>
<td>-2.15</td>
<td>14.58%</td>
<td>25.48%</td>
<td>0.7444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>40.49%</td>
<td>52.81%</td>
<td>5.7128</td>
</tr>
<tr>
<td>guitars</td>
<td>0.252</td>
<td>-3.11</td>
<td>15.34%</td>
<td>26.35%</td>
<td>0.2339</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>37.94%</td>
<td>50.11%</td>
<td>2.3745</td>
</tr>
<tr>
<td>houses</td>
<td>0.218</td>
<td>-2.62</td>
<td>16.09%</td>
<td>27.16%</td>
<td>0.2511</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>44.51%</td>
<td>56.88%</td>
<td>3.4209</td>
</tr>
<tr>
<td>leaves</td>
<td>0.470</td>
<td>-6.08</td>
<td>29.43%</td>
<td>41.44%</td>
<td>0.8699</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>46.60%</td>
<td>59.28%</td>
<td>3.0674</td>
</tr>
<tr>
<td>motorbikes_side</td>
<td>0.181</td>
<td>-2.34</td>
<td>15.63%</td>
<td>27.64%</td>
<td>0.2947</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>38.62%</td>
<td>51.64%</td>
<td>3.7305</td>
</tr>
</tbody>
</table>
Learning and Chernoff Information (9)

Figure: Filtering in the Caltech101 database, $T = T_{\text{min}}$ vs $T = 0$
Following the Chernoff theorem which exploits the Sanov’s theorem (quantifying probability of rare events), for \( n \) i.i.d. samples distributed by following \( Q \), the probability of error for the test with hypotheses \( Q = p_{on} \) and \( Q = p_{off} \) is given by:

\[
pe = \pi_{on} 2^{-nD(p_{\lambda}||p_{on})} + \pi_{off} 2^{-nD(p_{\lambda}||p_{off})}
\]

\[
= 2^{-n \min\{D(p_{\lambda}||p_{on}), D(p_{\lambda}||p_{off})\}}
\]

being \( \pi_{on} \) and \( \pi_{off} \) the priors.

Choosing \( \lambda \) so that \( D(p_{\lambda}||p_{on}) = D(p_{\lambda}||p_{off}) = C(p_{on}, p_{off}) \) we have that Chernoff information is the best achievable exponent in the Bayesian probability of error.
Figure: Filtering results in three environments with $T = T_{min}$ and $T = 0$
MD-Scale Saliency (Motivation)

Figure: Several example images from the *Bristol* dataset (RGB representation of 31 spectral bands).
Let $X$ be a $d$-dimensional random variable, and $f(x)$ its pdf. Let $A = \{A_j| j = 1, \ldots, m\}$ be a partition of $X$ for which $A_i \cap A_j = \emptyset$ if $i \neq j$ and $\bigcup_j A_j = X$. Then, we have [Stowell& Plumbley, 09]:

$$f_{A_j} = \frac{\int_{A_j} f(x)}{\mu(A_j)} \quad \hat{f}_{A_j}(x) = \frac{n_j}{n \mu(A_j)},$$

where $f_{A_j}$ approximates $f(x)$ in each cell, $\mu(A_j)$ is the $d$-dimensional volume of $A_j$. If $f(x)$ is unknown and we are given a set of samples $X = \{x_1, \ldots, x_n\}$ from it, being $x_i \in \mathbb{R}^d$, we can approximate the probability of $f(x)$ in each cell as $p_j = n_j/n$, where $n_j$ is the number of samples in cell $A_j$. 
Differential Shannon entropy is then asymptotically approximated by

\[ \hat{H} = \sum_{j=1}^{m} \frac{n_j}{n} \log \left( \frac{n}{n_j} \mu(A_j) \right), \]

and such approximation relies on the way of building the partition. It is created recursively following the data splitting method of the k-d tree algorithm. At each level, data is split at the median along one axis. Then, data splitting is recursively applied to each subspace until an uniformity stop criterion is satisfied.
The aim of this stop criterion is to ensure that there is an uniform density in each cell in order to best approximate $f(x)$. The chosen uniformity test is fast and depends on the median. The distribution of the median of the samples in $A_j$ tends to a normal distribution that can be standardized as:

$$Z_j = \sqrt{n_j} \frac{2med_d(A_j) - \min_d(A_j) - \max_d(A_j)}{\max_d(A_j) - \min_d(A_j)}$$

When $|Z_j| > 1.96$ (the 95% confidence threshold of a standard Gaussian) declare significant deviation from uniformity. Not applied until there are less than $\sqrt{n}$ data points in each partition.
KD-Partitions and Divergence

KDP Total-Variation Divergence

The total variation distance [Denuit and Bellegem, 01] between two probability measures \( P \) and \( Q \) for a finite alphabet, is given by:

\[
\delta(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|.
\]

Then, the divergence is simply formulated as:

\[
\delta(P, Q) = \frac{1}{2} \sum_{j=1}^{p} |p_j - q_j| \in [0, 1], \quad p(A_j) = \frac{n_{x,j}}{n_x} = p_j p(A_j) = \frac{n_{o,j}}{n_o} = q_j
\]

where \( p_i \) and \( p_j \) are the proportion of samples of \( P \) and \( Q \) in cell \( A_j \).
KD-Partitions and Divergence

Figure: KDP divergence. Left: $\delta = 0.24$. Right: $\delta = 0.92$
Multi-Dimensional Saliency Algorithm

Alg. 2: MD Kadir and Brady scale saliency algorithm

**Input:** $m-$dimensional image $I$, initial scale $s_{\text{min}}$, final scale $s_{\text{max}}$

**for each pixel $x$ do**

**for each scale $s_i$ between $s_{\text{min}}$ and $s_{\text{max}}$ do**

1. Create a $m-$dimensional sample set $X_i = \{x_i\}$ from $\mathcal{N}(s_i, x)$ in image $I$;
2. Apply kd-partition to $X$ in order to estimate
   \[
   \hat{H}(s_i, x) = \sum_{j=1}^{m} \frac{n_j}{n} \log \left( \frac{n}{n_j} \mu(A_j) \right)
   \]
3. If $i > s_{\text{min}} + 1$ then
   - If $\hat{H}(s_{i-2}, x) < \hat{H}(s_{i-1}, x) > \hat{H}(s_i, x)$ then
     - Compute Divergence: $W = \delta(X_{i-1}, X_{i-2}) = \frac{1}{2} \sum_{j=1}^{r} |p_{i-1} - p_{i-2}|$
     - Entropy weighting $Y(s_{i-1}, x) = \hat{H}(s_{i-1}, x) \cdot W$
   - Else
     - $Y(s_{i-1}, x) = 0$
4. $Y(s_{i-1}, x) = 0$

**end**

**end**

**Output:** An array $Y$ containing weighted entropy values for all pixels on image at each scale
KD-Partitions and Divergence

Figure: KNNG vs KDP Scale Saliency time from 1 to 31 bands. Top: $s_{\text{min}} = 5$, $s_{\text{max}} = 8$ (64 bins). Bottom row: $s_{\text{min}} = 5$, $s_{\text{max}} = 20$ (32 bins).
Figure: Deviation from the theoretical entropy of uniform (in the range $[-3, 3]^d$) and Gaussian (zero mean and $\Sigma = I$) distributions of the entropy estimated by the k-d partition method (KDP) and by the Leonenko et al. estimator for different values of $k$ ($K = 2 \ldots 5$).
Figure: The number of features decreases when increasing dimension and preserving the range of scale.
## MD-(Gabor) Saliency for Textures

### KDP Total-Variation Divergence

- Use Brodatz dataset (111 textures and 9 images per category: 999 images).
- Use 15 Gabor filters for obtaining multi-dimensional data.
- Both graylevel saliency and MD saliency are tuned to obtain 150 salient points.
- Use each image in the database as query image.
- Use: saliency with only RIFT, only spin images, and combining RIFT and spin images.
- Retrieval-recall results strongly influenced by the type of descriptor used.
MD-(Gabor) Saliency for Textures (2)

Figure: Salient pixels in textures. Left: MD-KPD. Right: Graylevel saliency
MD-(Gabor) Saliency for Textures

Figure: Average recall vs number of retrievals.
Conclusions & Future

- We propose an novel multi-dimensional saliency method.
- Applications: Hyperspectral images (keypoints or band selection) and Texture Categorization.
- KD-partitions estimators very efficient though not so consistent as Leonenko’s one. However we need relative estimations not precise ones.
- The number of entropy peaks decreases when increasing dimensions. It seems convenient to find the mos informative features in advance. For instance what is the optimal filter bank for textures? (future).
- In a MD context we may apply the filtering approach proposed in our previous work to speed-up categorization.
References


References (4)
