Entropy Estimation and Multi-Dimensional Scale Saliency

P. Suau, F. Escolano
Robot Vision Group
Universidad de Alicante, Spain
{*pablo,sco*}*@dccia.ua.es*

Abstract

In this paper we survey two multi-dimensional Scale Saliency approaches based on graphs and the k-d partition algorithm. In the latter case we introduce a new divergence metric and we show experimentally its suitability. We also show an application of multi-dimensional Scale Saliency to texture discrimination. We demonstrate that the use of multi-dimensional data can improve the performance of texture retrieval based on feature extraction.

1. Introduction

High level vision tasks usually rely on the results provided by image processing or feature extraction algorithms. The interest regions detected by feature extraction algorithms should satisfy several properties: they must be informative, distinguishable and invariant to a wide range of transformations[1]. The work in this paper is focused on the Scale Saliency algorithm by Kadir and Brady [2]. This algorithm is theoretically sound, due to the fact that it uses Information Theory in order to search the most informative regions on the image. Although its poor performance for matching problems [6], it has been shown to perform well in image categorization tasks [5]. Furthermore, it has been successfully applied before to this kind of problems [8][7].

The Scale Saliency algorithm [2] detects salient or unpredictable regions on an image. Shannon's entropy is used to measure the saliency of an image region. Given a pixel x, its entropy at scale s is computed from the grayscale intensity pdf of the circular region R_x of radius s, centered over x. The intensity pdf is approximated by means of an intensity histogram where $P_{d,s,x}$ is the probability that the intensity value $d \in D$ is found in R_x (in the case of a grayscale image, $D = \{0, \ldots, 255\}$). The algorithm works as follows: firstly, entropy is estimated for all pixels x in the image, using all scales s in a range of scales between s_{min} and s_{max} (Eq. 1). Next, entropy peaks (local maxima in scale space) are selected (Eq. 2). Then, entropy peaks are weighted by means of a self-dissimilarity metric between scales (Eq. 3). The aim of this step is to reinforce salient features that were detected at its characteristic scale. Finally, a subset of the salient features is selected, in order of weighted entropy (Eq. 4). These selected features are the most salient features of the image.

$$H(s,x) = \sum_{d \in D} P_{d,s,x} \log_2 P_{d,s,x} \quad (1)$$

$$S = \{s : H(s-1,x) < H(s,x) > H(s+1,x)\} \quad (2)$$

$$W(s,x) = \frac{s^2}{2s-1} \sum_{d \in D} |P_{d,s,x} - P_{d,s-1,x}| \quad (3)$$

$$Y(s,x) = H(s,x)W(s,x) \quad (4)$$

The application of the algorithm summarized above to higher dimensional data is straightforward. For instance, in RGB color images, where each pixel is assigned three different intensity values (corresponding to the three RGB channels), the local intensity pdf may be estimated from a 3D histogram. In general, for nD data, the same algorithm can be applied using a nD histogram for entropy and self-dissimilarity computation.

Two problems arise from this extension to the multi-dimensional domain, due to the curse of dimensionality. Firstly, the complexity order of the algorithm exponentially increases with data dimensionality. And secondly, higher dimensional data yields sparser histograms, that are less informative. These issues make the use of the original Scale Saliency algorithm unfeasible in the case of $n \geq 4$ dimensions.

This paper extends our previous work in [10]. We summarize two different modifications of the Scale Saliency algorithm in order to apply it to the multi-

[1] Several authors prefer the term *covariant*, referring to image features that adapt to the transformation applied to the image.

dimensional domain. Both versions of this Multi-dimensional Scale Saliency algorithm (MDSS) are based on alternative entropy and self-dissimilarity (divergence between scales) estimators. Therefore, no histograms are used and we may overcome the previously stated limitations of the original algorithm. The main contributions of this paper are: firstly, we propose a new divergence based on data partition, and we experimentally demonstrate its suitability. Secondly, we analyze both MDSS approaches in order to select the most appropriate one for feature extraction tasks. Finally, we show an example of application to texture categorization.

2. MDSS based on K-Nearest Neighbour Graphs

Firstly, we present a MDSS approach based on graphs. In this approach, each pixel $x_i \in X$ is represented as a d-dimensional vector. The neighbourhood R_x of a pixel is represented by an undirected and fully connected graph $G = (V, E)$, being the nodes $v_i \in V$ the d-dimensional vectors representing $x_i \in R_x$ and E the set of edges connecting each pair of nodes. The weight of each edge is the Euclidean distance in $\mathcal{R}^d$ between its two incident nodes. Entropy and divergence are estimated from the K-Nearest Neighbour Graph (KNNG), a subset of the fully connected graph, that connects each node to its k neighbours. In the case of entropy estimation, we apply the method proposed by Kozachenko and Leonenko [3]:

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} \log\left((N-1)e^{-\psi(k)} B_d (\rho_{k,N-1}^{(i)})^d\right) \tag{5}$$

were $|V| = N$, B_d is the volume of the d-dimensional unit ball, $\rho_{k,N-1}^{(i)}$ is the k-nearest neighbour of i when taking the rest of $N-1$ samples, and $\psi(z)$ is the digamma function.

In the case of self-dissimilarity between scales, we propose to apply the Friedman-Rafsky test. Let s be the scale in which an entropy peak has been found. In order to weight that entropy value, we must calculate the dissimilarity with respect to scale $s-1$. Let X_s and X_{s-1} be the set of nodes of R_x at scales s and $s-1$. Since $X_{s-1} \subset X_s$ (new pixels are added to the previous ones as we increase the scale), the test only requires to build the KNNG from X_s and to count the amount of edges in this KNNG that connect a node from X_s/X_{s-1} to a node from X_{s-1}. One minus this number of edges is a consistent estimator of the Henze and Penrose divergence.

3. MDSS based on the k-d partition algorithm

The second MDSS approach is based on the k-d partition algorithm by Stowell *et al.* [9]. As in the approach presented above, each pixel in R_x is represented as a d-dimensional vector. The d-dimensional feature space is recursively split into cells following the data splitting method of the k-d tree algorithm. At each level, data is split by their sample median along one axis. Then, data splitting is applied to each subspace until an uniformity stop criterion is reached. The aim of this stop criterion is to produce cells with uniform empirical distribution, in order to best approximate the underlying pdf. The data partition yields a set $A = \{A_j\}$ of p cells, and then entropy estimation is given by

$$\hat{H} = \sum_{j=1}^{p} \frac{n_j}{n} \log\left(\frac{n}{n_j}\mu(A_j)\right) \tag{6}$$

where $\mu(A_j)$ is the volume of the cell A_j, n_j is the number of samples in A_j and n is the the total number of samples in R_x.

Regarding the self-dissimilarity between scales, we propose a new divergence metric inspired by the k-d partition algorithm. Our k-d partition based divergence metric follows the spirit of the total variation distance [1], but may also be interpreted as a L1-norm distance. The total variation distance between two probability measures P and Q in the case of a finite alphabet is given by

$$\delta(P, Q) = \frac{1}{2}\sum_{x} |P(x) - Q(x)| \tag{7}$$

Let $f(x)$ and $g(x)$ be two distributions, from which we gather a set X of n_x samples and a set O of n_o samples, respectively. If we apply the partition scheme of the k-d partition algorithm to the set of samples $X \bigcup O$, the result is a partition A of $X \bigcup O$, being $A = \{A_j | j = 1, \ldots, p\}$. In the case of $f(x)$, the probability of any cell A_j is given by

$$p(A_j) = \frac{n_{x,j}}{n_x} = p_j \tag{8}$$

where $n_{x,j}$ is the number of samples from X in cell A_j. Conversely, in the case of $g(x)$ the probability of each cell A_j is given by

$$q(A_j) = \frac{n_{o,j}}{n_o} = q_j \tag{9}$$

where $n_{o,j}$ is the number of samples from X in the cell A_j. Since both sample sets share the same partition A, and considering the set of cells A_j a finite al-

phabet, we can compute the total variation distance between $f(x)$ and $g(x)$ as

$$D(O||X) = \frac{1}{2}\sum_{j=1}^{p} |p_j - q_j| \qquad (10)$$

The latter distance metric can be use as a self-dissimilarity measure in the Scale Saliency algorithm, since it satisfies $0 \leq D(O||X) \leq 1$. The minimum value $D(O||X) = 0$ is obtained when all the cells A_j contain the same proportion of samples from X and O. By the other hand, the maximum value $D(O||X) = 1$ is obtained when all the samples in any cell A_j were gathered from the same distribution.

4. Experimental results

In this section we introduce additional experiments to those shown in our previous work [10]. Firstly we test the validity of our k-d partition based divergence. We also present an application of our algorithm to the texture categorization problem.

The experiments in [10] were aimed to compare the computational time of the MDSS algorithms and the quality of the extracted features. In the former case we demonstrated that the computational order decreased from exponential (due to the use of histograms in the original Kadir and Brady algorithm) to linear with respect to data dimensionality. The computational efficiency of the k-d partition approach is remarkably higher when compared to the rest of algorithmsIn the case of the quality of the extracted features, we applied a repeatability test in order to check the stability of the extracted features over a wide range of transformations, using the image dataset provided by Mikolajczyk *et al.* [6]. The results showed that none of the MDSS approaches performs better than the other one in all circumstances. From these experiments (and others not included here due to the lack of space), and although both approaches reported similar results in terms of repeatability we concluded that the k-d partition based approach is far superior to the graph based approach.

4.1. Divergence validation

In order to validate our k-d partition based divergence, we compare its trend with that of the Friedman-Rafsky test: we compare the divergence of two sample sets gathered from two Gaussian distributions, starting with the same mean and variance, as we increase the distance between Gaussian centers until the probability that the samples overlap is low. The results for different data dimensionalities d are shown in Fig 1

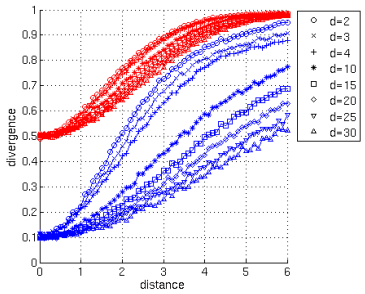

Figure 1. Divergence metrics comparison.

(Friedman-Rafsky test in red, k-d partition divergence in blue). In the case of both tests, the divergence (y axis) increases with the distance between Gaussian centers (x axis). The values or Friedman-Rafsky test lie in the range $[0.5, 1]$. The range of values in the case of our k-d partition divergence depends on data dimensionality, but it is generally wider and follows the trend of the Friedman-Rafsky test.

5. An application: texture discrimination

In this section we show an application of our algorithm to the texture discrimination method of Lazebnik *et al.* [4]. In this method, texture images are represented by means of signatures $S = \{(t_1, w_1), \ldots, (t_n, w_n)\}$, where t_i is a texton and w_i is its relative weight. Firstly, affine features are extracted from grayscale information on the image, and a descriptor is computed for each one. This descriptor may be a Rotation Invariant Feature Transform descriptor (RIFT) and/or a spin images descriptor (for a complete description of both descriptors, see [4]). Agglomerative clustering is applied to all descriptors in an individual image in order to build its signature. The textons are the center of these clusters, and their relative weight is computed as the number of descriptors in a cluster divided by the total number of descriptors in the image. Signatures are compared by means of the Earth Mover's Distance (EMD – see [4] for more detail). Our approach uses multi-dimensional Scale Saliency for feature extraction. Features are extracted from 15D data, computed after applying a Gabor filter bank (consisting of 15 Gabor filters with different orientations and wavelengths) to all pixels on the image.

In Fig. 2 we show the results of our texture retrieval experiment. In this experiment, that shows the performance of a given texture representation, all images in the Brodatz dataset[2] are used as query image once. For

[2] http://www.ux.uis.no/~tranden/brodatz.html

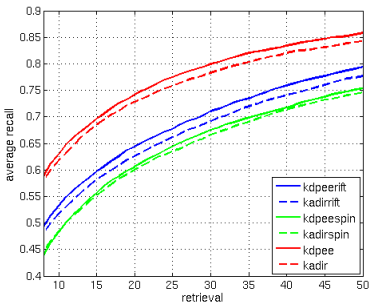

Figure 2. Results of the texture categorization experiment.

each image query, we select images from the database in increasing order of EMD. The result is a plot that shows the average recall of all query images (being recall the number of images from the class of the query image retrieved so far divided by the total number of images in that class) versus the number of closest images retrieved. In Fig. 2 we compare the performance of the grayscale Scale Saliency and k-d partition based multi-dimensional Scale Saliency methods using only RIFT (*kadirrift* and *kdpeerift*, respectively), only spin images (*kadirspin* and *kdpeespin*), and combining RIFT and spin images (*kadir* and *kdpee*). In order to combine RIFT and spin images in the retrieval task, the total distance between two images is computed adding the normalized EMDs estimated for each individual descriptor.

Multi-dimensional data increased the performance of the texture retrieval task for each tested descriptor. However, its impact is not as noticeable as the impact of choosing an adequate descriptor. As can be seen in Fig. 2, the average retrieval is strongly affected by this last factor. The worst results are obtained for the case of spin images. RIFT increases the average recall, but the most significative improvement is achieved when combining both.

6. Conclusions and future work

The Scale Saliency algorithm by Kadir and Brady can be naturally extended to process multi-dimensional data. However, its computational efficiency remarkably decreases with data dimensionality. We survey two approaches of multi-dimensional Scale Saliency based on alternative entropy and divergence metrics which computational order is linear with respect to data dimensionality. Our analysis show that the k-d partition approach should be preferred over the graph based approach. We introduced a new divergence metric based on the k-d partition algorithm and the total variation distance, and we experimentally showed its suitability. Finally, we showed a practical application of our approach in the context of texture representation.

Our future work is addressed to evaluate the application of multi-dimensional data in other Computer Vision problems, like video processing or image retrieval. In the texture categorization context, we should also study the impact of using different Gabor filter banks, or even different input data. This is a combinatorial problem that may be treated with machine learning methods like feature selection.

7. Acknowlegdements

This work has been partially funded by the project TIN2008-04416 of the Spanish Government.

References

[1] M. Denuit and S. V. Bellegem. On the stop-loss and total variation distances between random sums. *Statistics and Probability Letters*, 53:153–165, 2001.

[2] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[3] L. Kozachenko and N. Leonenko. On statistical estimation of entropy of a random vector. *Problems of Information Transmission*, 23:95–101, 1987.

[4] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1265–1278, 2005.

[5] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 2, pages 1792–1799, 2005.

[6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[7] P. Newman and K. Ho. Slam-loop closing with visually salient features. In *proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 635–642, 2005.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[9] D. Stowell and M. D. Plumbley. Fast multidimensional entropy estimation by k-d partitioning. *IEEE Signal Processing Letters*, 16(6):537–540, 2009.

[10] P. Suau and F. Escolano. A new feasible approach to multi-dimensional scale saliency. In *proceedings of the 11th International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 77–88, 2009.