# Some Observations and Ideas About the Shape Index Descriptor

Kim Steenstrup Pedersen

March 26, 2014

This document collects some observations and thoughts about the shape index descriptor.

## 1 Histograms versus Kernel Density Estimators

Pedersen et al [1] defines the shape index descriptor using smooth histogram estimators of the distribution of a feature $f : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$ having magnitude or weight $F : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$,

$$H(f_i; \mathbf{r}_0) = \int_{\mathbb{R}^2} F(\mathbf{r}; \sigma) A(\mathbf{r}; \mathbf{r}_0, \alpha) B(f_i, \mathbf{r}; f, \beta) \, \mathrm{d}\mathbf{r} \quad , \tag{1}$$

with $\mathbf{r} = (x, y)^T$ and where $f_i$ denotes the histogram binning variable and will act as the bin center for a specific choice of binning aperture function $B$. If we have $N$ bins then we can think of the histogram $\{H(f_i) | i = 1, \ldots, N\}$ as an $N$ dimensional vector, $\mathbf{H} \in \mathbb{R}^N_+$.

The function $A$ localizes the descriptor to specific parts of the image indicated by $\mathbf{r}_0$ and can in principle have any form, but we choose a Gaussian aperture function of scale $\alpha$

$$A(\mathbf{r}; \mathbf{r}_0, \alpha) = \frac{1}{2\pi\alpha^2} \exp\left( -\frac{(\mathbf{r} - \mathbf{r}_0)^2}{2\alpha^2} \right) . \tag{2}$$

The binning function could be the Gaussian function of $\beta$ bin scale (what Koenderink refers to as tonal scale)

$$B(f, \mathbf{r}; f_i, \beta, \sigma) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left( -\frac{(f(\mathbf{r}; \sigma) - f_i)^2}{2\beta^2} \right) , \tag{3}$$

or have any other form we like.

Pedersen et al [1] fixes the bin centers $f_i$ to tile the feature domain equidistantly. The specific choice of number of bins $N$, locations of $f_i$ and bin scale $\beta$ is an open question. We can potentially by-pass the first two by rephrasing the problem in terms of kernel density estimation.

In kernel density estimation we position a kernel function at each data point, when evaluating the density at a specific location (not necessarily at fixed locations or at data points) we let neighboring data points vote for the density at the location. Assuming we have a data set of $M$ features measured at different locations

We can formulate the kernel density estimator in the continuous domain similar to our formulation for histograms,

$$H(\hat{f}; \mathbf{r}_0) = \int_{\mathbb{R}} \int_{\mathbb{R}^2} F(\mathbf{r}; \sigma) A(\mathbf{r} - \mathbf{r}_0; \alpha) B(\hat{f} - f(\mathbf{r}); \beta) \, \mathrm{d}\mathbf{r} \mathrm{d}f \quad , \tag{4}$$

For this to work as a descriptor we need to be able to compare descriptors and we would like it to have a minimal memory footprint for storage. For comparison we could simply use standard metrics for comparing probability densities such as Kullback- Leibler divergence, mutual information, etc. However, having to store all $N$ data points in order to evaluate the density is not appealing. Instead we could sample the continuous function $H(\hat{f})$ at fixed locations $f_i$ and store the densities at these locations thereby forming a vector representation, $\mathbf{H} \in \mathbb{R}_+^N$.

An alternative to sampling the continuous function $H(\hat{f})$ is to represent the function by its moments. From the moments we can reconstruct the distribution via its characteristic function. However, this leaves the question of which and how many moments do we need to faithfully represent the distribution.

The issue of selecting bin scale $\beta$ still remains open and one can consider moving away from selecting the same scale for all bins and instead locally in the feature domain optimize the scale. For both the histogram and kernel density estimators this is important. Choosing to small a scale $\beta$ will lead to overfitting in both cases. For the histogram estimator the bin scale is thightly coupled to the number of bins $N$ — i.e. lowering $\beta$ should be followed by an increase of $N$ else we will have an estimate with holes of poor density estimates between bin centers. When increasing $\beta$ both estimators delivers smooth density estimates. In fact the bin kernel $B$ and scale $\beta$ forms a tonal scale-space (see the work on locally orderless images by Koenderink [2]). There is a functional relationship between the tonal scale $\beta$ and the aperture scale $\alpha$ and measurement scale $\sigma$. For discrete images the aperture scale $\alpha$ defines how many pixels are to be included in the density estimate. For small $\alpha$ we will have a small set of pixels which will force us to increase the bin scale $\beta$ and for the histogram estimator the number of bins $N$ in order to avoid overfitting to the data set. As such $\beta$ and $N$ has an inverse proportional relationship to $\alpha$. The measurement scale affects the range of possible feature values $f$ and therefore also has an effect on the location of data points in the feature domain as part of the kernel density estimator and the binning in the histogram estimator. However, this relationship is not easy to express in the general setting.

TODO: Make an example illustrating the difference between the histogram estimator and the kernel density estimator.

## 2 Scale selection using entropy

Inspired by the work of Sporring and Weickert [3] lets consider using generalized entropy as a measure of feature strength and scale selection criteria.

Generalized entropy or Renyi entropy is a spectrum of entropy measures for a discrete probability distribution $\mathbf{p}$ defined as

$$S_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \log \sum_{i=1}^{N} p_i^\alpha \tag{5}$$

when $\alpha \neq 1$ and for $\alpha = 1$ we get the ordinary Shannon-Wiener entropy

$$S_1(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i \ . \tag{6}$$

The parameter $\alpha$ defines a spectrum of entropy measures. By varying the generalized entropy parameter $\alpha$ we can put different weighting on high and low probability bins in the histogram.

Information-rich histogram features should contain some structure to make them distinct. A histogram representing a uniform distribution does not provide much information. Our scale space histogram features undergoes a smoothing as we increase the scale $\sigma$ (or $\beta$ for that matter) which leads to an increase in entropy towards maximal entropy when the histogram represents a uniform distribution. We can use entropy measures to single out scales at which the histogram obtains a local minima in entropy, which corresponds to situations where the histogram has a particular "peaked" structure.

An alternative approach is to follow the idea proposed by Sporring and Weickert [3] where they consider the rate of change of entropy across scale as an indication of interesting scales,

$$c_\alpha(\mathbf{p}(\sigma)) = \frac{\partial S_\alpha(\mathbf{p}(\sigma))}{\partial \sigma} \ . \tag{7}$$

In practice this can simply be implemented with differencing of entropy across scale.

Interesting scales could either be those scales which leads to the smallest rate of change $c_\alpha(\mathbf{p}(\sigma))$ corresponding to scale ranges where the entropy of the histogram $\mathbf{p}(\sigma)$ is stable / a plateau, or those scales which leads to the largest rate of change. The latter corresponds to scale ranges where the entropy changes the most. Hence we can identify interesting scales by finding the extrema of $c_\alpha(\mathbf{p}(\sigma))$ across scale.

Entropy measures provide a summary of the information captured in the histogram and the entropy curves across scale provides a fingerprint of the histograms behaviour across scale.

The question remains how to select $\alpha$. We can put emphasis on the contribution of high probability areas by choosing $\alpha >> 1$. If we instead want to focus on low probability areas we should choose $\alpha < 0$.

There are different ways we can use this scale selection principle in a descriptor. We can use it to select one or more scales and then build a single scale descriptor at each of these scales. We can also use the identified scales as either defining the base scale in a multi-scale descriptor or defining the scale levels in a multi-scale descriptor. For the latter one has to consider how to fix the number of selected scales (i.e. fixing the dimensionality of the feature vector).

# References

[1] K. S. Pedersen, K. Stensbo-Smidt, A. Zirm, and C. Igel, "Shape index descriptors applied to texture-based galaxy analysis," in *Proceedings of ICCV'13*, 2013.

[2] J. J. Koenderink and A. J. van Doorn, "The structure of locally orderless images," *IJCV*, vol. 31, no. 2/3, pp. 159–168, 1999.

[3] J. Sporring and J. Weickert, "Information measures in scale-spaces," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 1051–1058, April 1999.