A Trainable Low-level Feature Detector

Peter Hall, Martin Owen, and John Collomosse
Department of Computer Science
University of Bath
pmh | cspmjo | jpc@cs.bath.ac.uk

Abstract

We introduce a trainable system that simultaneously filters and classifies low-level features into types specified by the user. The system operates over full colour images, and outputs a vector at each pixel indicating the probability that the pixel belongs to each feature type. We explain how common features such as edge, corner, and ridge can all be detected within a single framework, and how we combine these detectors using simple probability theory. We show its efficacy, using stereo-matching as an example.

1. Introduction

Low-level feature detection has long been of major interest to Computer Vision. Attention has usually been directed toward developing *prescriptive* methods; that is, methods that provide an objective definition of the feature to be detected based on some prior, usually tacit, model. Typical features include edges [2], corners [6], and (to a lesser extent) ridges [10]. Such systems have many advantages, but a disadvantage is that they tend to disagree with humans observation. For example, a typical edge map will both admit and omit edges when compared to an average of human observations.

We aim to detect features that more closely agree with the average of human observations, and so introduce a novel system to detect low-level features. Ours is not a prescriptive system, but is instead user trained. The user simply points to examples of the their chosen features. Given a full colour image as input, the system outputs a probability vector at each pixel, each component of which is the probability that the pixel belongs to the corresponding feature type.

The importance of a user trained system is that detected features tend to be more salient, when compared to prescriptive methods. The importance of a multi-classifier is similar; some applications require edges, others corners, many both. Our multi-classifier allows a user to choose features types to detect, and combine them in simple but well principled ways. Our system is very flexible, it has the potential to be used in many applications, such as stereo-matching, subimage matching, segmentation, and modelling.

The fact we train from user input places our system in the same general area as that of Konishi *et al* [9], who provide a trainable system for statistical edge detection. We differ in three ways. First we provide a multi-classifier rather than a dichotomiser. Second we use circular sampling rather than traditional "Cartesian" sampling, which gives the advantage of certain invariants. Finally we introduce novel terms for "visibility", "coherence", and "rarity", claiming that a feature must be visible, coherent, and rare to be of interest. Although heuristic in nature, these latter add considerably to the efficacy of our classifier; in particular their combination can be regarded as some measure of feature salience.

Our feature detector is based on circular sampling. Smith and Brady have already argued in favour of circular sampling [1]. Krüger and Felsberg [4] use radial-polar coordinates to compute the "intrinsic dimension" of features, which enables them to classify features as flat, edge, or corner; the method is prescriptive, they make no provision for other possible feature types. The reason we favour circular sampling is that it allows a characterisation which is invariant to orientation and reflection. Our approach is also invariant to colour shifts and colour inversion, and is robust to colour scaling (caused by luminance changes, say). These invariances mean that features of a given type, edges say, all map to the same point in the parameter space we use, thus giving a strong signal. This is to be preferred over Cartesian based characterisations in which features tend to be more thinly distributed in parameter space; edges, for example, would lie on a ring-like manifold.

2. A trainable low-level feature detector

We use a colour image as input, in RGB format. At a pixel $\mathbf{x} = (x, y)^T$ we consider coloured samples $\mathbf{c}(\rho, \theta)$ for discrete values of radius ρ and angle θ. The sampling disc has N_ρ rings and N_θ "spokes". We assume that up to a local radius ρ_{max} the feature can be classified a single type, such

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

as "flat", "edge", "corner", or "ridge"; clearly ρ_{max} defines the scale of the feature.

We can now characterise the feature at $\mathbf{x}$ and radius ρ by considering the magnitude of the differential signal

$$u(\theta) = \left| \frac{\partial \mathbf{c}(\rho, \theta)}{\partial \theta} \right| \tag{1}$$

which we compute using central differences based on the Euclidean distance between RGB colours. We parameterise this signal using

$$f(\omega) = \frac{\mathcal{F}[u(\theta)]}{(\sum_\theta u(\theta)^2)^{1/2}} \tag{2}$$

$$f(\omega) \leftarrow f(\omega) \setminus f(0) \tag{3}$$

where $\mathcal{F}$ is the Fourier transform. Using $|\mathcal{F}|$ ensure orientation and mirror invariance. Normalising by unit power is needed to enhance discrimination between some feature types; it raises contrast. Removal of the "dc" term, $f(0)$, removes dependence on mean luminance. Zernike moments [12] were considered an alternative for feature characterisation, we have implemented Zernike moments but found no advantage.

During training users choose as many classes as they wish, and then point to example features in real images to train the system. The system requires users to point to "unknown" examples too. If the user chooses just one class — edges, perhaps — our system thus becomes a dichotomiser. In any case it builds a parametric description, we use Gaussian Mixture Models (GMMs), of the training sample distribution for each class, thus enabling us to estimate $p(f|k)$ as the class conditional probability that feature type k is responsible for observed feature f. We use the ratio of samples in each class to obtain a prior $p(k)$, thus enabling an estimate of the posterior probability $p(k|f)$ using Bayes theorem.

The term just introduced is useful in characterising the *type* of feature, but suffers because normalisation boosts noise. We suppress noise using a measure based on simple model of just-noticeable-differences (jnds): given a colour $\mathbf{c}$ its jnd is a distance τ such that no colours in the closed ball $[\mathbf{c}, \tau]$ can be discriminated from $\mathbf{c}$, and all colours in the complement of the closed ball $[\mathbf{c}, \tau]$ can be so discriminated. We determined the threshold τ for random noise — we asked users to adjust τ until a colour square looked just flat. We used RGB colour space, arguing that the distance function $\tau(\mathbf{c})$ is sufficient for perceptually based colour measures, leaving us free to choose any convenient colour space. We use the full sample disc $\mathbf{c}(\rho, \theta)$ to estimate the probability a feature rises above noise level, and define the *perceptual gradient magnitude*, $\phi(\rho, \theta)$, via the Laplacian:

$$\phi(\rho, \theta) = \frac{1}{\tau(\mathbf{c})} \left(\sum_{i=1}^{3} \left(\frac{\partial \mathbf{c}_i}{\partial \rho} \right)^2 + \left(\frac{1}{\rho} \frac{\partial \mathbf{c}_i}{\partial \theta} \right)^2 \right)^{1/2} \tag{4}$$

$$\bar{\phi} = \frac{1}{N_\rho} \sum_{\rho=1}^{\rho_{\max}} \frac{1}{\rho} \frac{1}{N_\theta} \sum_\theta \phi(\rho, \theta) \tag{5}$$

where $\mathbf{c}_i$ is the ith colour channel. The radially-weighted average, $\bar{\phi}$, is a crude measure of visibility. We asked users to dichotomise training samples in visible / non-visible examples, and so empirically modelled the probability of visibility as

$$p_v(\bar{\phi}) = (1 + \mathrm{erf}((\bar{\phi} - \mu)/\sigma))/2; \tag{6}$$

The constants were chosen experimentally as $\mu = 0.6$ and $\sigma = 0.4$. both of these are in jnd units. One might expect $\mu = 1$, that is one jnd. The lower value is in part a consequence of the $1/rho$ weight, but may also imply that structure in a feature somehow boosts effective contrast compared to unstructured noise (which we used to determine one jnd).

To further eliminate pixels which appear at some level to be a feature but are actually part of a chaotic area of the image, we introduce a coherence measure. We use this in preference to entropy, say, as used by Kadir and Brady [8]. Our measure is based on the idea that any part of a coherent region is similar to any other part. We use the Euclidean distance between every pair of rings, up to some scale $\rho_{\max}$, and compute the probability that a pixel is coherent as follows

$$d(\rho_1, \rho_2) = \left(\sum_\theta |\mathbf{c}(\rho_1, \theta) - \mathbf{c}(\rho_2, \theta)|^2 \right)^{1/2} \tag{7}$$

$$D = \sum_{(\rho_1, \rho_2) \in \mathcal{R}} d(\rho_1, \rho_2) \tag{8}$$

$$p_c(D) = \frac{1}{2} \left(1 - \mathrm{erf} \left(\frac{D - \nu}{\lambda} \right) \right) \tag{9}$$

where $\mathcal{R}$ is the set of all distinct Cartesian pairs (ρ_1, ρ_2) in the sample disc. The constants, $\nu = 0.25$ and $\lambda = 0.2$, in the error function were chosen by observation over many images of differing kinds.

The terms introduced so far are picture independent. We propose a final term that is picture dependent — *rarity* which others have associated with salience [13]. We estimate the probability of salience by first estimating the likelihood for every full colour sample window in the image. Given a likelihood measure, $\pi(w)$, for each window w we define the probability of rarity as the fraction of windows that are less rare:

$$p_r(w) = \frac{|\{\pi(w') < \pi(w), w', w \in W\}|}{|\{\pi(w) \in W\}|} \tag{10}$$

where W is the domain of valid windows (do not over lap image border).

Since $p(k|f)$, $p_v(\bar{\phi})$, $p_c(D)$, and $p_r(w)$ all depend on location, we define the probability that a pixel is salient and of type k as

$$p(\hat{k}|\mathbf{x}) \quad = \quad p(k|f)p_v(\bar{\phi})p_c(D)p_r(w) \qquad (11)$$

The only exception to this is if the type chosen is "flat", meaning the user has decided that areas of flat colour are of interest. In this special case we use $1 - p_v$, the probability of not being visible, in place of p_v. In any case $p(\hat{k}|\mathbf{x})$ indicates the probability that pixel $\mathbf{x}$ belongs to "interest class" k. The sum of such probability is unity, at every pixel. The components in the "multi-spectral" output image, $p(\hat{k}|\mathbf{x})$, can be combined in anyway that conforms with standard probability theory. For example if "edge" and "corner" are defined classes we can consider $p(\text{edge}|\mathbf{x})$ as an edge map, $p(\text{corner}|\mathbf{x})$ as a corner map, and $p(\text{edge}|\mathbf{x}) + p(\text{corner}|\mathbf{x})$ as a combined corner-edge map.

The above approach classifies artifacts at a constant ρ, and so at a single, constant scale. However classification can vary over scale. For example, an artifact classified as an edge at small scales might be classified a ridge at larger scales. Whilst this allows for a scale-space description there are contexts where a definitive classification decision needs to be made taking a range of scales into account. We have found that simply averaging $p(\hat{k}|\mathbf{x})$ over several scales (typically $\rho_{\max} = 1$ to 6, inclusive) gives an effective a scale-independent description.

3. Results

We now illustrate with qualitative results and present and quantitative support for the claim that features specified by our classifier are often more useful for stereo matching than features obtained using a popular prescriptive method. We trained a type classifier using a fixed radius of 3 pixels using the types "flat", "edge", "ridge", "corner", and "unclassified". We used 4 images (two faces, one indoor still life, one outdoor of a car against trees). As Figure 1 demonstrates, training at a fixed scale in no way prevents us from classifying over a range of scales, although there may be benefit to scale-specific training. Simple averaging over the scale range can lead to improved results, as shown in the figure. This provides features that persist over scale, which is said to be a condition of salience [11]. We show output from a typical Canny edge detector and a typical Harris filter for visual comparison. The photograph in Figure 1 was not in the training set, and typical of the output we obtained for many test images of many kinds (faces, cars, outdoors, indoors). See Figure 2 for and example of the classifier run on a different image.

We now turn to quantitative results. Essential for good stereo matching are points which are distinctive and salient.

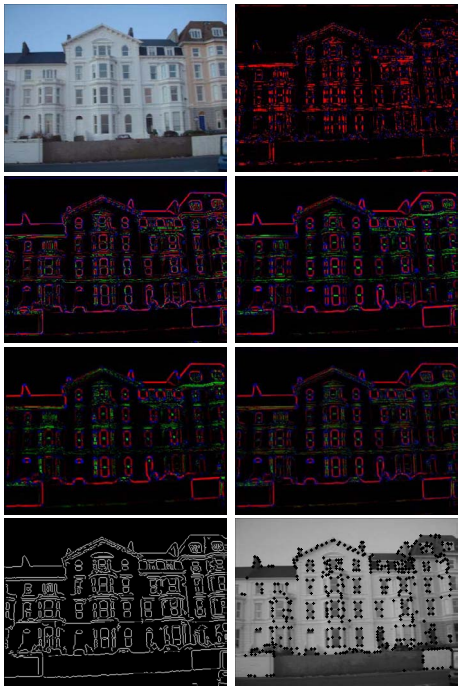

Figure 1. A photograph (top left) classified at scales 1 to 4 inclusive, reading top-left to bottom-right. Showing edges (red), ridges(green), and corners(blue), flat or unclassified points in black. Average of scaless 1–6 in the third row, right–hand column. Typical Canny edges (bottom-left) and Harris features (bottom-right) from the intensity image shown for comparison.

Figure 2. Classification of a still life image.

They need to be strongly associated with important image features to ensure the presence of reliable matches in the second image. It is for these reasons we propose the classifier corner data is aptly suitable for feature matching. We use the "corner" image from $p(\hat{k})$, thresholded at 0.5, this threshold is justified because it picks out points which are most likely to be corners. Such corner points tend to form clusters, from each cluster we chose the corner with the highest probability. These points were used in stereo matching over a series of image pairs and compared to Harris corners [6] as this is a popular tool. The Harris detector relies

upon parameters, we used a single set of parameters which was optimal over a range of images. The images ranged from simple block-worlds to complicated outdoor images that included trees and other foliage.

Prospective matches between the images were made using a least square difference between the feature neighbourhoods. From these two sets of pairs of potentially matching feature pairs, homographies were calculated using RANSAC [5]. The relative quality of the different feature sets could thus be compared by the proportion of the matching pairs which are deemed inliers by RANSAC. When using RANSAC the number of iterations made depends on the expected number of outliers. Knowing in advance that the classifier points are less likely to be outliers means a fewer number of iterations need be made to ensure a sample is made without any outliers. Hartley and Zisserman [7] give the number of iterations N as

$$N = \frac{\log(1-p)}{\log(1-(1-\epsilon)^s)} \tag{12}$$

where p is the desired probability of making a sample without any outliers. Here $p = 0.99$, s is the number of points taken in each sample, $s = 4$ and ϵ is the probability of a random sample being an outlier. For Harris corners the mean proportion of outliers was $\epsilon = 0.7471$, resulting in $N = 1123$. For our classifier $\epsilon = 0.6340$, giving $N = 254$: a small difference in performance can lead to a dramatic improvement in efficiency.

4. Conclusion

We have introduced a user-trainable classifier. It is improved by the introduction of novel terms for visibility, rarity, and coherence. We have in part addressed the issue of scale by averaging over scales. The need for visibility, coherence, and rarity may cause a slight concern on the grounds that such terms should be integrated into the classifier, not least because it enables a fully user-trainable system. Yet close examination of the joint population distribution $p(\mathrm{type}, \mathrm{visibilit\ y}\mathrm{coherence}, \mathrm{rarit\ y})$ shows it to contain sharp discontinuities; separating the terms out as we have makes modelling much simpler. We are in the process of examining non-parametric approaches for describing the joint distribution, and associated marginals. As it stands, the system has some analogy to weak classifiers: none of the classifiers is alone sufficient, yet their combination proves to be so.

Another concern might exist over the way we handle scale-independent decisions. Currently we simply average over many scales, but initial work on user training over many scales indicates an advantage, particularly eliminated "grazing edges", which pass through the edge of the sample disc without reaching the centre — these can be mistaken for corners. The system does not localise as well as prescriptive methods. For example it also can produce "patches", as around corners — but such patches tend to occur where corners are curved, and in such cases it is not always clear what is meant by "corner".

To its advantage, our approach is flexible, because users can define whatever features they wish. Furthermore, it is easy to extract a single feature map, or combine feature maps. It produces useful features as our quantitative results in stereo-matching show. Elsewhere we have used it in the production of novel non-photorealistic images: it enabled us to render different classes of features in different ways [3]. We believe our detector will find many more applications in which to be useful. The fact is produces probability maps might make it useful as part of a larger learning system.

References

[1] S. S. . J. Brady. Susan – a new approach to low level image processing. Technical report, FMBIB, Oxford University, 1995.

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 8(6):34–43, June 1986.

[3] J. P. Collomosse and P. M. Hall. Genetic painting: A salience adaptive relaxation technique for painterly rendering. Technical Report CSBU 2003-04, University of Bath, October 2003.

[4] N. K. . M. Felsberg. A continuous formulation of intrinsic dimension. In *Proceedings British Machine Vision Conference*, pages 260–270. BMVA, 2003.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with apphcatlons to image analysis and automated cartography. *ACM Computing Surveys (CSUR)*, 14:3–71, March 1982.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 189–192, Manchester, UK, 1998.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge University Press, 2000.

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

[9] S. Konishi, A. L. Yuille, J. M. Caughlan, and S. C. Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 25(1):57–74, January 2003.

[10] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1996.

[11] D. Marr and E. Hildreth. Theory of edge detection. *Proc. of Royal Soc. of London B*, 207:187–217, 1980.

[12] M. Teague. Image analysis via general theory of moments. *Journal of Optical society of America*, 70(8):920–930, 1979.

[13] K. Walker, T. Cootes, and C. Taylor. Locating saliernt object features. In *Proceedings British Machine Vision Conference*, volume 2, pages 557–567. BMVA, 1998.