On Representing Edge Structure for Model Matching

T.F. Cootes, C.J. Taylor
Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester M13 9PT U.K.
t.cootes@man.ac.uk

Abstract

We show how a novel, non-linear representation of edge structure can be used to improve the performance of model matching algorithms and object verification/recognition tasks. Rather than represent the image structure using intensity values or gradients, we use a measure which indicates the orientation of structures at each pixel, together with an indication of how reliable the orientation estimate is. Orientations in flat, noisy regions tend to be penalised whereas those near strong edges are favoured. We demonstrate that this representation leads to more accurate and reliable matching between models and new images, and leads to better recognition/verification of faces in an access control task. [1]

1. Introduction

When building models of the appearance of objects it is advantageous to choose a representation of the image structure which can capture the features of interest in a way that allows a reliable comparison between model and image, and is invariant to the sorts of global transformation that may occur. For instance, when building statistical appearance models [2, 4] or eigen-face models [12] it is common to represent the structure by a vector of intensity values, normalised by a linear transform so as to be invariant to global changes in brightness and contrast. By sampling across the whole region, all image structures are represented and all pixels treated equally (though the statistical analysis with then typically favour pixels in some regions over others, as dictated by the data). However, models based on raw intensity tend to be sensitive to changes in conditions (such as lighting and camera parameters). Thus models built on one data set may not perform well on images taken under different conditions. This is often the case in face interpretation work, in which models built from images taken under controlled conditions in the lab. do not work well on the more noisy and unconstrained images one tends to find in the real world.

Edge based representations tend to be less sensitive to lighting conditions than raw intensity measures. Thus an obvious alternative to modelling the intensity values directly is to record the local image gradient in each direction at each pixel. Although this yields more information at each pixel, and at first glance might seem to favour edge regions over flatter regions, it is only a linear transformation of the original intensity data. Where model building involves applying a linear PCA to the samples, the resulting model will be almost identical to one built from raw intensities, apart from some effects around the border, where computing the gradients includes some background information into the model.

In this paper we present a new representation. Rather than just recording the intensities at each pixel, we compute a measure which represents the local orientation of structures at each pixel, together with an indication of how reliable the orientation estimate is. In particular we use a non-linear normalisation function which is chosen so as to accentuate likely edges and suppress structure with is likely to be noise.

In the 2D greyscale case we end up with a texture vector which contains two values for each pixel in the modelled region, which can be used to compare two regions or a model instance and an image region.

Statistical models of shape and texture can be matched to new images using the Active Appearance Model algorithm (AAM) [2, 4]. We demonstrate how the new representation can be used to improve the performance of the AAM, leading to more accurate and reliable model matching. We also show how the representation can lead to better results in a facial recognition task than a pure intensity based model.

2. Background

Eigen-faces [12] model the statistics of the intensities in a region of an image, and have been widely used for object

[1] This appears in Proc CVPR 2001, Volume 1, pp.1114-1119

location and recognition. Moghaddam and Pentland generalised this to include models of smoothed canny edge images [11].

Edge maps are widely used in the image registration community [5]. However, they tend to use either linearly normalised gradients or squared gradients, or non-maximally suppressed edges (all pixels other than those thought to be exactly on the edge are set to zero).

Edge orientation images have been used for face recognition by Hond and Spacek [6], who created histograms of edge orientation over regions and obtained good results. Our method can be thought of as a weighted version of edge orientation, in which strong edges are given more weight than weak edges. Rather than use a histogram we model the edge structure at every pixel.

Bosch *et.al.* [1] have used a non-linear normalisation step on intensities in echo-cardiograms as pre-processing before building an appearance model. The intention is to modify the strongly non-gaussian noise statistics of such images into more normal shaped distributions. This gives significantly improved results. However, this was on intensities, not edge statistics, and the approach described may not be optimal when applied to edge statistics.

Glasbey and Mardia [5] use a Fourier von-Mises model of the distribution of image values to compare two images in a non-linear registration problem. They propose a comparison function constructed in the fourier domain, which generalises a wide range of common measures. The computation in the fourier domain used above makes it harder to apply to non-square regions than the image based method proposed below. Also, in the above *frequencies* with low amplitude are penalised, whereas in our method *pixels* with low edge strength are penalised, potentially a more local effect.

Kittler *et.al.* [7] demonstrated that different types of image normalisation could have a significant effect on a face verification task. They found that histogram normalisation tended to perform well over a range of experiments.

3. Statistical Models of Appearance

An appearance model can represent both the shape and texture variability seen in a training set. The training set consists of labelled images, where key landmark points are marked on each example object. The training set is usually labelled manually, though automatic methods are being developed.

These can be thought of as generalisations of eigen-patches or eigen-faces [12, 11] in which, rather than represent a rigid region, we model the shape of the region and allow it to deform.

Given such a set we can generate a statistical models of shape and texture variation (see [2, 4] for details). The shape of an object can be represented as a vector $\mathbf{x}$ and the texture (grey-levels or colour values) represented as a vector $\mathbf{g}$. The appearance model has parameters, $\mathbf{c}$, controlling the shape and texture according to

$$\begin{array}{rcl} \mathbf{x} & = & \bar{\mathbf{x}} + \mathbf{Q}_s\mathbf{c} \\ \mathbf{g} & = & \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c} \end{array} \tag{1}$$

where $\bar{\mathbf{x}}$ is the mean shape, $\bar{\mathbf{g}}$ the mean texture and $\mathbf{Q}_s$,$\mathbf{Q}_s$ are matrices describing the modes of variation derived from the training set.

An example image can be synthesised for a given $\mathbf{c}$ by generating a texture image from the vector $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$ (see Figure 2 below).

Such a model can be matched to a new image, given an initial approximation to the position, using the Active Appearance Model algorithm [3]. This uses a fast linear update scheme to modify the model parameters to minimise the difference between a synthesized image and the target image.

Appearance models and AAMs have been shown to be powerful tools for face image interpretation [2] and medical image interpretation [10, 1].

These appearance models typically represent the region of interest using linearly normalised intensity values. However, this tends to be sensitive to lighting changes and differences between cameras (both important problems in face recognition scenarios). A more stable representation is likely to be either an edge gradient, or an edge orientation measure. In the following we will describe a novel method of representing edge orientation for statistical appearance models.

4. Representing Orientation

We wish to represent the orientation of structure at each pixel in a region. The simplest method of measuring orientation is to compute the local gradients, g_x and g_y, at the point and set $\theta = \tan^{-1}(g_y/g_x)$. However, because the angles can 'wrap around', θ is not a useful value to record if we wish to compare two angles quickly. To get around this, a common approach is to represent the angle using points on a unit circle, $(\cos\theta, \sin\theta) = (g_x/g, g_y/g)$ where $g = \sqrt{g_x^2 + g_y^2}$. Angles can then be compared simply using Euclidean distance (for alternatives, see Mardia [8]).

We propose to represent the image structure using a measure which captures the orientation of image structure measured at each pixel, together with a measure of our belief in the accuracy of the measurement. Essentially this will be strong when we are on a strong edge, as there the orientation is well defined, but will be weak or zero in essentially flat noisy regions of the image. Below we will describe the approach, showing how the orientation can be represented either modulo 2π, or modulo π (useful in the cases in which the polarity of the edge can change).

In the 2D case, we perform the following at each pixel of the region of interest:

- Estimate the local gradients in x and y, g_x, g_y using a suitable filter
- Compute the magnitude of the gradient, $g = \sqrt{g_x^2 + g_y^2}$
- Apply a non-linear function, $f(g)$ to obtain a representation of edge direction and strength, $(g'_x, g'_y) = f(g)(g_x/g, g_y/g)$

The normalising function $f(g)$ is chosen so as to accentuate likely edges and suppress structure with is likely to be noise. Without loss of generality, $0 \leq f(g) \leq 1$ for all g.

For example, an effective function has been found to be $f(g) = |g|/(|g| + g_0)$, where g_0 is the mean or median of the expected values of g. In the case of the appearance model, this can be the mean or median of g measured over the current region being represented. This function has the desired property that values of g less than g_0 tend to get mapped towards zero and those above g_0 (those more likely to be edges of interest) get mapped towards 1.

An alternative is $f(g) = P_n(g)$, where $P_n(x)$ is the cumulative probability distribution of the edge strengths due to effective noise in the image. Again, values of g likely to be due to noise effects get mapped toward zero, those less likely to be noise are treated as interesting structure and are effectively enhanced. Note that if the noise on each pixel is gaussian, the noise on the derivatives, g_x,g_y, is also gaussian and thus $g^2 = g_x^2 + g_y^2$ will be distributed as χ^2 - allowing an analytic estimate of $P_n(g)$ in terms of gamma functions. An estimate of the noise variance can be obtained from the distribution of g over the sampled region, or can be precomputed for the image as a whole, as appropriate.

4.1. Orientation modulo π

In some cases it is advantageous to represent the orientation of the edge structure only up to modulo π, essentially identifying the direction of the edge gradient but not its polarity. This is useful in cases where we do not know the background in advance - the object may be darker than the background or lighter than it, so we only know the position and direction of edges, but not their polarity (dark to light or light to dark). This can be dealt with by first representing the edge direction in polar co-ordinates, $(g_x, g_y) \rightarrow (g, \theta)$. We then double the angle, and map back; $(h_x, h_y) = (g\cos 2\theta, g\sin 2\theta) = (g_x^2 - g_y^2, 2g_xg_y)/g$. We then apply a suitable non-linear normalisation as described above,

$$\begin{aligned}(g'_x, g'_y) &= f(g)(h_x/g, h_y/g) \\ &= f(g)(g_x^2 - g_y^2, 2g_xg_y)/g^2\end{aligned} \tag{2}$$

4.2. Edge Structure Appearance Models

To build an appearance model, the normalised texture vector $\mathbf{g}$, which would typically be a concatenation of all the intensity values, is replaced with a vector twice as long, of the form $\mathbf{g}^T = (\ldots, g'_{xi}, g'_{yi}, \ldots)$. All other calculations remain the same. When searching, the same approach is used when sampling the image. Rather than compare the normalised image texture vector with that generated by the model, we compare the non-linearly normalised gradient vectors.

The effect of this is that the matching process seeks to manipulate the model so as to match the edge structure as closely as possible. Noisy, approximately flat regions tend to be represented as near zero values and thus have little effect on the optimisation. Experiments demonstrate this leads to more accurate and reliable location of the position of the boundaries of the structures of interest, particularly when the target image has significantly different lighting conditions compared to the original training set.

Choice of Reference Frame

Note that the directions along which the gradients are measured, and their scaling, should be defined in the model frame, to ensure that the calculation is independent of the rotation and scale of the model instance in the image.

4.3. Face Appearance Model

We trained an Edge Structure Appearance Model on 100 images taken at random from a set containing people with a wide range of expressions (Figure 1). 68 points were used to represent the shape of each face. 28 shape parameters were required to explain 98% of the shape variation. 5000 pixels were use to represent the image texture. 48 combined appearance model parameters were required to represent 95% of the combined variation. Figure 2 shows the effect of varying the first two model parameters (for display purposes, the intensity is proportional to the y-component of the edge orientation).

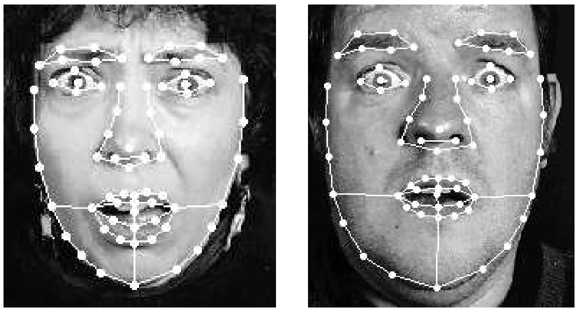

Figure 1. Examples of faces from the expression set

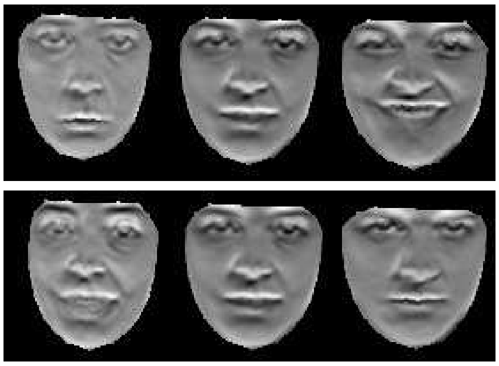

Figure 2. Effect of varying first two facial appearance model parameters, c_1,c_2 by ± 3 standard deviations from the mean.

5. Experiments

5.1. Search Performance

We used the representation to build appearance models of faces, and used the Active Appearance Model algorithm to search new images. We adopted the following procedure to test the search performance: On each test image we start a search with the mean model shape at positions displaced from the known optimal centre by ± 10 pixels in x and y. After running a multi-resolution AAM we compute the mean error between model points and hand labelled points (pt-pt), and the mean distance of the model points to the appropriate hand labelled boundaries (pt-crv).

Performance on similar images

We performed the above test with the edge structure appearance model on 100 different images drawn from the same data set as the training set (thus they were new images of the same people). We obtained mean point-point errors of 1.7 pixels and point-curve errors of 1.2 pixels.

For comparison, we trained an intensity appearance model on the same data. This required 26 combined parameters to explain 95% of the total variation. Repeating the search experiment gave almost identical results to the edge structure model.

Performance on different data set

The experiment was repeated on a different data set. The same model (that trained on an expression varying set) was tested on 100 face images from the XM2VTS dataset [9]. These were of different people in completely different lighting conditions, taken with a different camera.

In addition to the edge structure (ES) and normalised intensity (NI) models, we built and tested models for normalised gradient (NG), in which we record the linearly normalised x and y gradients at each pixel and a modified edge structure (ES2) model in which we use the angle modulo π representation described above.

Table 5.1 summarises the results. Successful convergence was declared if the mean point to point error was less than 10 pixels. Figure 3 shows the distributions of point-curve errors for the four cases. The edge structure models both give better search results than the intensity or gradient models.

Model Type	Success	Pt-Pt Error	Pt-Crv Error
ES	88%	7.0 ± 1.2	3.9 ± 1.1
ES (2)	89%	7.2 ± 1.2	4.1 ± 1.1
NI	61%	7.4 ± 1.5	4.2 ± 1.3
NG	79%	7.4 ± 1.4	4.3 ± 1.2

Table 1. Comparative performance of AAMs (trained on expression images) when searching unseen XM2VTS set images, starting at displaced pose

Edge structure (2) uses the direction modulo π. If $\mathbf{g}$ is the original gradient vector, it uses $(2 * g_x * g_y, g_x^2 - g_y^2)/(|\mathbf{g}|^2 + e_0)$.

Similar images, larger displacements

We repeated the experiment, training on 100 images from the XM2VTS data and testing on a different set of images from the same database. In this case we displaced by ± 20 in x and y, and ran a 3 level multi-resolution search. Results are given in table 5.1

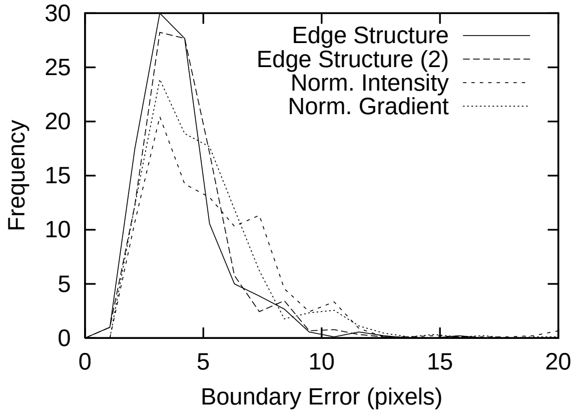

Figure 3. Distribution of boundary errors after AAM search on XM2VTS images (model built on expression set)

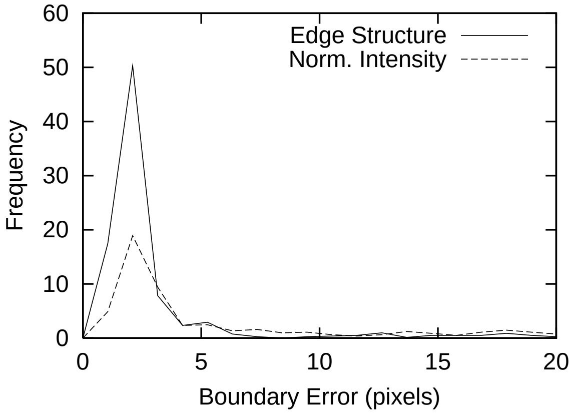

Figure 4. Distribution of boundary errors after AAM search on unseen XM2VTS images (model built on M2VTS)

Model Type	Success	Pt-Pt Error	Pt-Crv Error
ES	81%	4.0 ± 1.4	2.6 ± 0.9
NI	38%	4.5 ± 1.6	3.0 ± 1.0

Table 2. Comparative performance of AAMs when searching unseen XM2VTS set images, starting at displaced pose (model trained on XM2VTS)

Model Type	Success	Pt-Pt Error	Pt-Crv Error
ES	97.4%	9.2 ± 2.5	5.9 ± 1.9
NI	65.0%	9.6 ± 2.6	6.1 ± 1.8

Table 3. Comparative performance of rigid AAMs when searching unseen M2VTS set images, starting at displaced pose (model trained on XM2VTS)

In this case the edge structure model gives a considerably better result than the linearly normalised grey-level models. In particular it is much less likely to diverge to an incorrect result. The distributions of curve errors are given in Figure 4.

5.2. Eigen-face Models

If we set the number of shape model modes to zero, we obtain an 'eigen-face' model, in which all variation is represented by the texture model. We can still use the AAM algorithm to perform efficient matching. Given an initial estimate of position, it refines the position, scale and orientation to optimise the match (minimise a sum of squares difference).

We trained such models on the 100 expression set images, then used them to search the 100 XM2VTS images. Success was declared if the point-point error was less than 20 pixels. Note that an eigen-face model cannot accurately locate all the points, it can only find the best rigid transformation mapping the model to the image. Table 5.2 summarises the results - again the edge structure model converges significantly more frequently than the normalised intensity model.

6. Recognition

To test the utility of the approach in recognition or classification applications, the following experiment was performed. An appearance model was built from 100 images of 19 different people with different expressions (about 5 images of each individual). LDA was applied to estimate an 18 dimensional identity subspace. The model parameters for each of 300 test images of the same people were projected into this subspace.

Note that no search was performed - in each case we matching the model to hand marked points. This allows us to compare verification performance of the representation without being confounded by search errors. In the following we are interested in the relative performance, rather than the absolute performance (which could probably be improved by more careful choice of model settings).

Appearance parameter vectors for each image were compared with those from all other images. This was done by computing using a normalised correlation distance in the subspace, essentially the cosine of the angle between two vectors. This has been found to be a useful way of discriminating between faces [7]. By comparing the angles between vectors for images of the same person to those between vectors for different people, ROC curves were generated and equal error rates estimated. This simulates the access control scenario in which a single image is used to register each person, then all the other people attempt access as that person. Table 4 shows the results - the edge structure model gives significantly better performance than the normalised intensity model.

Model Type	Equal Error Rate	
	XM2VTS	Expression
Edge Structure	5.8%	13.3%
Normalised Intensity	18.5%	10.6%

Table 4. Equal error rates for classification of face images from two data sets (model built on expression set)

A similar experiment performed on the XM2VTS face data set [9] (using 100 images for training and 120 pairs of images of individuals for testing) showed no significant difference between performance of the representations with both giving equal errors of about 11%.

We performed a third experiment, in which we used the model built from 100 expression set images to test the equal error rate of the verification on the XM2VTS data. In this case the edge structure model performed slightly worse than the intensity model (Table 4).

It is suspected that this is because the lighting conditions and expressions were similar for all the faces in this set. The edge structure approach appears to improve performance on sets where there is a lot of expression or lighting variation.

We are currently exploring this further.

7. Discussion and Conclusions

We have proposed a novel representation of edge structure which encodes both the orientation and our belief in the accuracy of the measurement. This tends to highlight strong edges and suppress responses in noisy areas of the image.

The representation can be used in statistical appearance models, both in rigid 'eigen-patch' type models and in deformable models which represent both shape and texture variation. We have shown that such a representation can lead to more reliable and accurate search compared to a model of normalised intensities, or linearly normalised edge gradient. The improvement is particularly noticable when training a model on one data set and testing it on another, taken under different conditions. This appears to be because edge orientation tends to be more robust to external factors such as lighting changes than linearly normalised intensities.

Where there is a significant lighting or expression variability in a face data set, the edge structure appearance model appears to give significantly improved results compared to a normalised intensity model. However, on data sets in which the images are taken under similar lighting conditions, with neutral expressions, there is no significant improvement, and there can be a degradation. We are currently investigating this further.

The improvements to the robustness and accuracy of model matching achieved with the representation suggest that it will be useful in many applications.

References

[1] H. Bosch, S.C.Mitchell, P.F.Boudewijn, P.F.Leieveldt, F.Nijland, O. Kamp, M. Sonka, and J. Reiber. Active appearance-motion models for endocardial contour detection in time sequences of echocardiograms. In *SPIE Medical Imaging*, Feb. 2001.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H.Burkhardt and B. Neumann, editors, 5^{th} *European Conference on Computer Vision*, volume 2, pages 484–498. Springer, Berlin, 1998.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, june 2001.

[4] G. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In 3^{rd} *International Conference on Automatic Face and Gesture Recognition 1998*, pages 300–305, Japan, 1998.

[5] C. Glaseby and K. Mardia. A penalized likelihood approach to image warping. *Journal of the Royal Statistical Society B*, 63(3), 2001.

[6] D. Hond and L. Spacek. Distinctive descriptions for face processing. In 8^{th} *British Machine Vison Conference*, volume 1, pages 320–329, Colchester, UK, 1997.

[7] J. Kittler, Y. Li, and J. Matas. On matching scores for lda-based face verification. In M. Mirmehdi and B. Thomas, editors, 11^{th} *British Machine Vison Conference*, volume 1, pages 42–51, Bristol, UK, Sept. 2000. BMVA Press.

[8] K. Mardia. *Statistics of Directional Data*. Academic Press, London, 1972.

[9] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. 2nd Conf. on Audio and Video-based Biometric Personal Verification*. Springer Verlag, 1999.

[10] S. Mitchell, P. Boudewijn, P.F.Lelievedt, R. van der Geest, H. Bosch, J. Reiber, and M. Sonka. Time continuous segmentation of cardiac mr image sequences using active appearance motion models. In *SPIE Medical Imaging*, Feb. 2001.

[11] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.