Multi-band Modelling of Appearance

Mikkel B. Stegmann, Rasmus Larsen

Informatics and Mathematical Modelling, Technical University of Denmark – DTU
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Email: {mbs, rl}@imm.dtu.dk

Abstract— **Earlier work has demonstrated generative models capable of synthesising near photo-realistic grey-scale images of objects. These models have been augmented with colour information, and recently with edge information. This paper extends the Active Appearance Model framework by modelling the appearance of both derived feature bands and an intensity band. As a special case of feature-band augmented appearance modelling we propose a dedicated representation with applications to face segmentation. The representation addresses a major problem within face recognition by lowering the sensitivity to lighting conditions. Results show that localisation accuracy of facial features is considerably increased using this appearance representation under diffuse and directional lighting and at multiple scales.**

Keywords— **Generative models, Active Appearance Models, Lighting Invariance, Face Recognition, Segmentation.**

I. Introduction

MODELS capable of synthesising complete images of objects have over the past few years proven their worth when interpreting unseen images. Applications include real-time tracking of deformable objects [1], [2], face recognition [3], [4], [5], and recovery of anatomical structures in magnetic resonance images [6], [7], [8], [9], ultrasound images [10] and x-rays [9], [11]. The key idea to all of these generative models is to perform a per-pixel comparison between unseen input images and synthesised images and subsequently drive these to equality.

In this paper, we investigate a generative model that has proven widely applicable. The Active Appearance Models (AAMs) [12], [13] have been applied to most of the examples given above. As Cootes et al. [14] the appearance of edge strength is modelled, but in contrast this is augmented with colour information and conventional raw intensities. We show that a considerable gain in accuracy can be achieved, merely by selecting a more appropriate representation of the particular object class being modelled. As such, this paper demonstrates that mature image processing methods can co-exist in rewarding symbiosis with a modern generative model-based vision technique.

II. Active Appearance Models

Active Appearance Models [12], [13] establish a compact parameterisation of object variability, as learned from a training set by estimating a set of latent variables. The modelled object properties are usually shape and pixel intensities. The latter is henceforward denoted *texture*. From these quantities new images similar to the training set can be generated.

Objects are defined by marking up each example with points of correspondence over the set either by hand, or by semi- to completely automated methods. The key to the compactness of these models lies in proper compensation of shape variability prior to modelling texture variability. Models failing in doing this, such as Eigen-faces [15], experience major difficulties in modelling variability in a compact manner.

Exploiting approximate prior knowledge about the local nature of the optimisation space, these models can be fitted to unseen images in a fraction of a second, given a reasonable initialisation.

Variability is modelled by means of a Principal Component Analysis (PCA), i.e. an eigen analysis of the dispersions of shape and texture. Shapes are brought into alignment using a Generalised Procrustes Analysis (GPA) [16], and textures are warped into correspondence using a thin-plate spline [17] or piece-wise affine warp, thereby compensating for any variation in shape. Let $\overline{\mathbf{x}}$ and $\overline{\mathbf{t}}$ denote the shape and texture mean, respectively. The (ranked) model parameters, $\mathbf{c}$, can then generate new instances in a simple linear manner:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{\Phi}_s \mathbf{c} \quad , \quad \mathbf{t} = \overline{\mathbf{t}} + \mathbf{\Phi}_t \mathbf{c} \tag{1}$$

where $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_t$ are eigenvectors obtained from the training set. The object instance, $(\mathbf{x}, \mathbf{t})$, is synthesised into an image by warping the pixel intensities of $\mathbf{t}$ into the geometry of the shape $\mathbf{x}$.

By defining a suitable measure of fit, $M(\mathbf{c}, \mathbf{I})$, the model could be matched to an unseen image, $\mathbf{I}$, using standard optimisation techniques such as conjugate-gradient, Levenberg-Marquardt or Metropolis-Hastings in a simulated annealing scheme. However, AAMs do not. Instead, residual vectors between the model and image, $\delta\mathbf{t} = \mathbf{t}_{model} - \mathbf{t}_{image}$ are regressed against known displacement vectors, $\delta\mathbf{c}$, using reduced-rank regression:

$$\delta\mathbf{c} = \mathbf{R}\delta\mathbf{t} \tag{2}$$

Embedded into an iterative updating scheme, this has proven to be a very efficient way of matching these models to unseen images. For large models (many texture samples) built on large training sets, this approach becomes quite resource demanding w.r.t. memory and computation. However, recent experiments [18] have shown that estimating the Jacobian, $\frac{\partial(\delta\mathbf{t})}{\partial\mathbf{c}}$, over the training set using a simple weighting scheme, in practice yields better results than the regression approach with far less computational and storage requirements. In the work below the regression approach has been taken.

This sums up the basic theory of AAMs. For further details refer to [13], [18], [19].

III. Multi-band AAMs

Contrary to the above univariate view upon images, the most frequently used image source – the RGB camera – is multivariate. Thus, collapsing the red, green and blue band into a single intensity band looses specificity. As Edwards et al. [12] we model multiple texture bands by simple concatenation. Any correlation between bands is to be picked up by the principal component analysis analogue to the recovered correlation along shape contours. The concept of *texture* is consequently extended to encapsulate any corresponding measurement over the training set. Let m denote the number of texture samples in band i:

$$\mathbf{u}_i = [\ u_{i1}\ u_{i2}\ \ldots\ u_{im}\ ] \tag{3}$$

The concatenated texture vector will then be for p texture bands:

$$\mathbf{t} = [\ \mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_p\ ] \tag{4}$$

Henceforth all AAM processing is left unchanged. This is multi-band modelling of appearance. As hinted this approach can be taken to all structures of corresponding input data, three dimensional problems, time-series [7], 3D+time etc. Often, the hard part is to obtain the correspondence, in particular for cases with sparse or incomplete data.

IV. The VHE representation

As a special case of multi-band appearance we propose a representation suitable for segmentation of face-like images. A secondary aim is to stress the ease with which one can add feature bands to create new representations suitable for a particular domain.

The statistical approach to model building has many striking advantages. Variability, dependencies, etc. are *learned* (estimated) from representative example solutions contrary to being *designed* (coded) explicitly into the model. However, some sources of variation are harder to generalise than others. Given a few people who smile it is a reasonable task to build a complete model of smiling mouths, i.e. a model that generalise well. This is due to the low intrinsic dimensionality of the geometrical deformation involved in a smile. On the contrary, lighting effects on a face are very hard to describe. The intrinsic dimensionality is high; 3D geometry of the face, skin surface (dry, sweaty), lighting (type, position, colour) etc.

As an alternative to learn the effects of lighting such as shadows and highlights, we propose a representation less sensitive to these. First, we notice that lighting effects have less influence on the hue band in the Hue, Saturation and Value (HSV) colour space. By modelling hue, we aim at obtaining the specificity of colour models without the sensitivity to effects of lighting. Second, as [14] we notice that edge estimators per se are less sensitive to lighting effects than raw intensities. Since edge estimators are implemented as numeric differential operators (e.g. Sobel-filters) these are unfortunately inherently sensitive to noise, which calls out for some degree of regularisation. This is often achieved through a modest filtering with a Gaussian kernel (preferably of the differential operator). Since this damps the high frequency content of an image, which is less desirable in a segmentation application, we choose to retain a pure intensity-based band. All together these three bands form the Value, Hue and Edge (VHE) representation:

- **V value** – The value (intensity) in the HSV colour-space.

- **H modified hue** – The angular hue, h, of an HSV representation modified to accommodate single-band storage. Since faces have little hue variation, the hue circle is here collapsed around the approximate circular mean, $\theta = 0$ and $\theta + \pi$ in the following way:

$$h_{mod} = \begin{cases} h & \text{if } h < \pi \\ 2\pi - h & \text{otherwise} \end{cases} \tag{5}$$

Though this introduces ambiguity in hues we expect this to be acceptable compared to the effects of wrapping angles.

- **E edge** – The edge strength, calculated as the gradient magnitude,

$$g = \sqrt{g_x^2 + g_y^2} \tag{6}$$

where g_x and g_y are horizontal and vertical gradient images obtained from numeric differential operators with a suitable amount of Gaussian smoothing.

V. Experiments

To test the hypotheses regarding the described VHE representation a database of 74 face images was compiled:

- **Set A** – 37 people facing front to the camera with a neutral facial expression. Lighting conditions were neutral using diffuse light from above.

- **Set B** – The same 37 people facing front to the camera with a new neutral facial expression. Partial non-diffuse lighting conditions were simulated by adding a directional light (horizontal lighting from the right, as seen from the camera).

Still images were recorded using a Sony DV video camera (DCR-TRV900E PAL) in 640×480 JPEG colour format and subsequently annotated using 58 landmarks. Refer to figure 1 for example images from Set A and B.

Grey-scale versions were obtained using the standard luminance-weighting scheme:

$$G = 0.30R + 0.59G + 0.11B \tag{7}$$

Alternatively, a Principal Component or Maximum Autocorrelation Factor transform [20] could be applied to the RGB bands to obtain grey-scale versions.

VHE versions of Set A and B were obtained using the procedure described previously. Refer to figure 3 for an example VHE transform. Notice the markedly lower horizontal resolution in the modified hue band. This is due

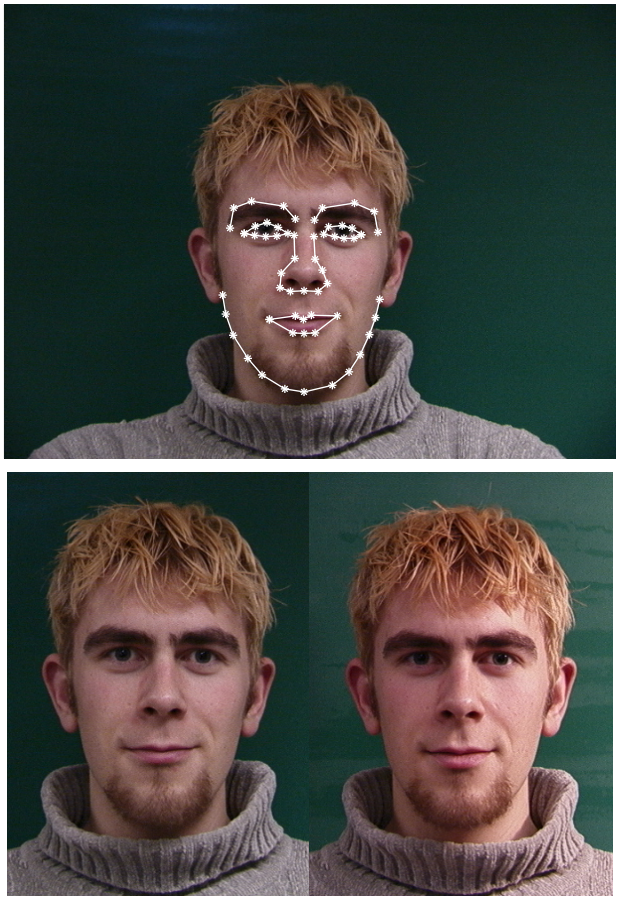

Fig. 1. Top row: Example annotation. Bottom row: Cropped example images from Set A using diffuse lighting (left) and Set B using directional lighting (right).

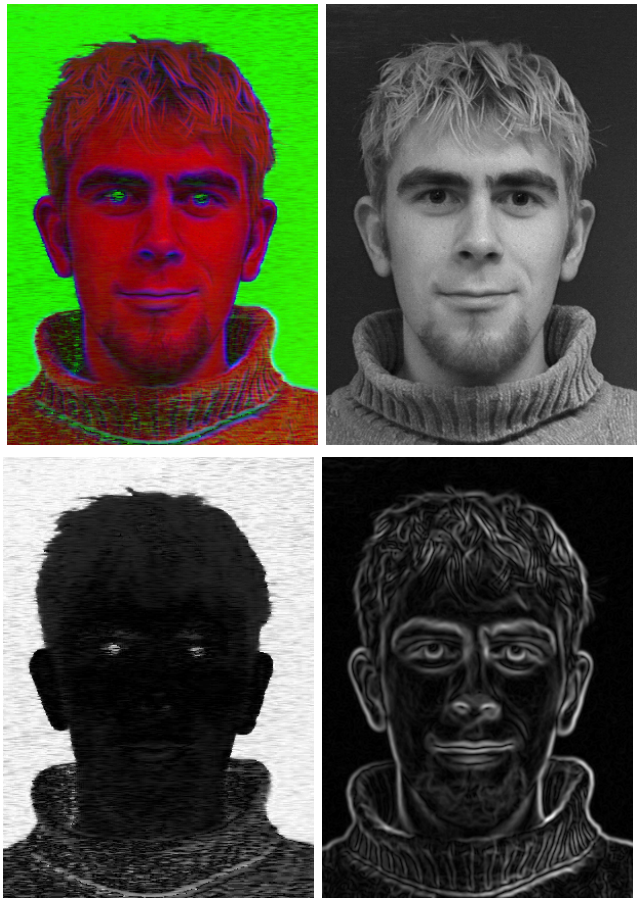

Fig. 3. Top row: VHE representation of a face (left) and value band (right). Bottom row: Modified hue band (left) and edge band (right).

Fig. 2. First combined principal mode, c_1, for an AAM built over the 37 images of Set A. Value, modified hue and edge bands are shown row-wise, top-down. The deformations are $c_1 = -3\sigma_1$ (left), mean (middle) and $c_1 = 3\sigma_1$ (right) where σ_1 is one standard deviation over the training set. Bands are stretched linearly for display.

TABLE I
LEAVE-ONE-OUT SEGMENTATION RESULTS.

	Mean pt.-pt.	Mean pt.-crv.
Grey-scale	2.73±0.78	1.35±0.46
Colour	2.84±0.75	1.35±0.40
VHE	2.63±0.64	1.27±0.40

TABLE II
SEGMENTATION RESULTS USING DIRECTIONAL LIGHTING.

	Mean pt.-pt.	Mean pt.-crv.
Grey-scale	3.51±0.85	1.78±0.49
Colour	3.22±0.67	1.67±0.44
VHE	2.91±0.65	1.40±0.36

Pt.-pt. measures Euclidean distance between corresponding landmarks of the model and the ground truth, whereas pt.-crv. measures the shortest distance to the curve in a neighbourhood of the corresponding ground truth landmark.

to the subsampling of the chrominance bands in the video formation and the JPEG compression scheme.[1]

In all experiments AAMs were initialised using a sparse global search exploiting the convergence distance of each parameter. Often this is only done in a few selected parameters. In this case position and scale were adequate. From the result of the global search a candidate set is chosen and iterated further until convergence. The best of these converged results denotes the initial position. To improve speed and robustness this is done on models built at multiple scales. For details see [21].

A. Segmentation of unknown identity using diffuse lighting

To assess the segmentation capabilities under standardised lighting conditions cross-validation were carried out on three different AAM representations of Set A: grey-scale, colour and VHE. To obtain optimal performance a leave-one-out scheme was used. Thus, 37 models were built from 36 examples each leading to 37 evaluations of each representation. Input images were subsampled to 320×240 pixels prior to any AAM processing. The texture models were ∼8000 pixels/band and it took on average 28 combined parameters to represent 95% of the variation observed in the training set.

The results in table I show a subtle increase in accuracy for the VHE representation compared to the standard grey-scale AAM. Unexpectedly, table I shows that the more specific colour AAM is slightly less accurate than the grey-scale AAM w.r.t. to mean pt.-pt. distance and equivalent for the mean pt.-crv. distance, though with a small decrease in uncertainty for the pt.-crv. measure. This indicates that the colour AAM slides more along contours. Though designed to handle changes in lighting, the VHE AAM seems to outperform both the grey-scale and colour AAMs by a modest amount under controlled lighting conditions.

B. Segmentation of known identity using directional lighting

Subsequently, the three representations were tested for their ability to segment known faces with subtle changes in expression but major changes in lighting. This was carried out by building three AAMs, grey-scale, colour and VHE on Set A (320×240 pixels). Refer to figure 2 for the first principal mode of the VHE AAM. All three models were subsequently tested on all images in Set B. Table II shows an increase in segmentation accuracy of 17% (pt.-pt.) and 21% (pt.-crv.) for VHE compared to Grey-scale. Further, the VHE has lower uncertainty estimates. From the error distributions in figure 4 it is noted that the VHE has a lower maximum error and in general a tail less heavy than the grey-scale and the colour AAM.[2]

[1]This step, which we had no control over, is motivated by a known decreased sensitivity in the human visual system to high-frequency chrominance content.

[2]The break-ups in the log-plot curves for the grey-scale and colour AAM are due to histogram bins with zero entries.

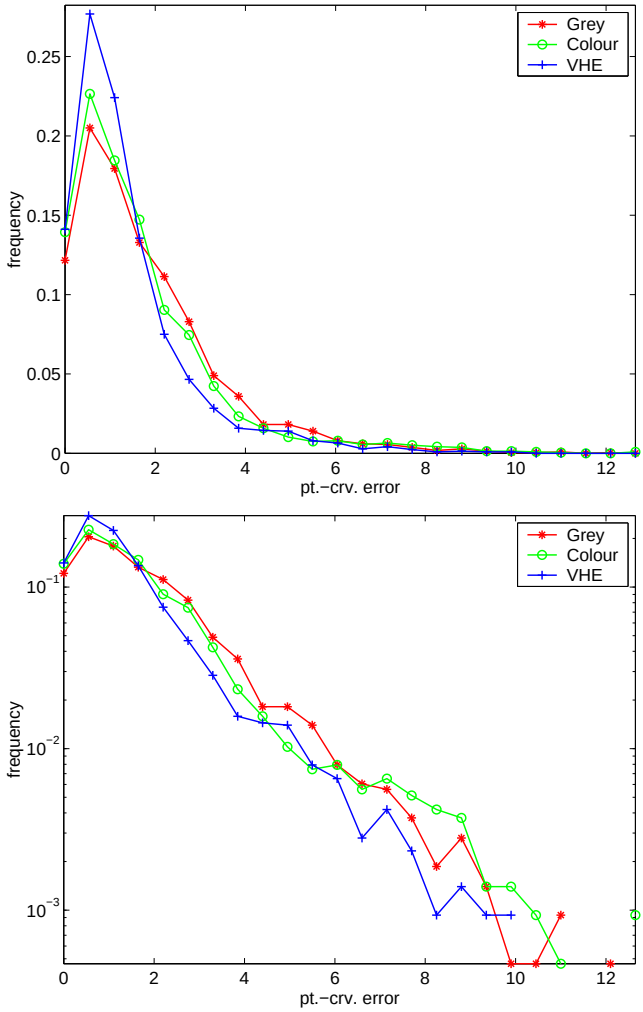

Fig. 4. Distribution of pt.-crv. errors using directional lighting. Shown as normal (top) and log (bottom) plots.

C. Accuracy at different scales

Occasionally, it is not feasible to build AAMs in the original input resolution. This can be due to constraints such as memory consumption, computation time etc. In a case with high-resolution input but a constraint on the model size one could ask whether a multi-band model should be chosen over a single-band model with a higher resolution.

To test this 18 AAMs were built. These were in six different resolutions using each of the three representations, grey-scale, colour and VHE. From Set A 27 examples were selected for training. Image resolutions spanned from 108×80 pixels to full input resolution at 640×480 pixels. The resulting model sizes were in the range 850 – 92118 texture samples. Using the described initialisation method all 18 models were evaluated against the remaining 10 examples of Set A. In figure 5 the mean pt.-pt. error is plotted against the model size. Here, pt.-pt. errors are measured in units of pixel width at the used resolution. While the VHE performs best, figure 5 stresses the fact that a simple pt.-pt. measure is worthless as performance indicator without the image resolution or model size given.

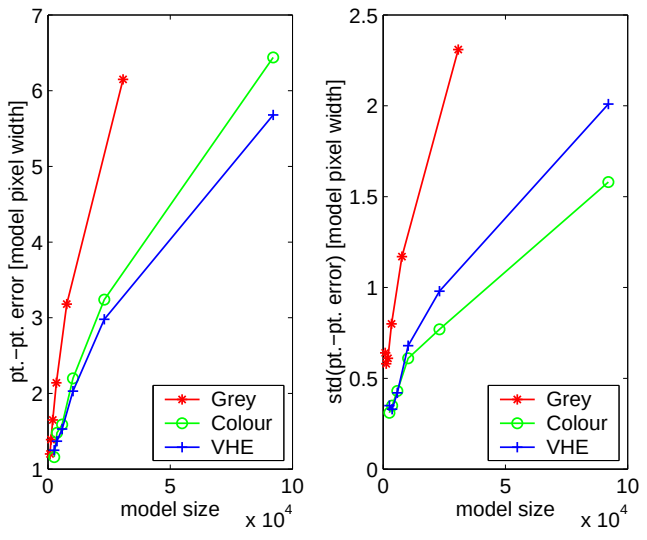

Fig. 5. AAM pt.-pt. accuracy measured as image pixel size vs. model size (left) and std. of image pixel size vs. model size.

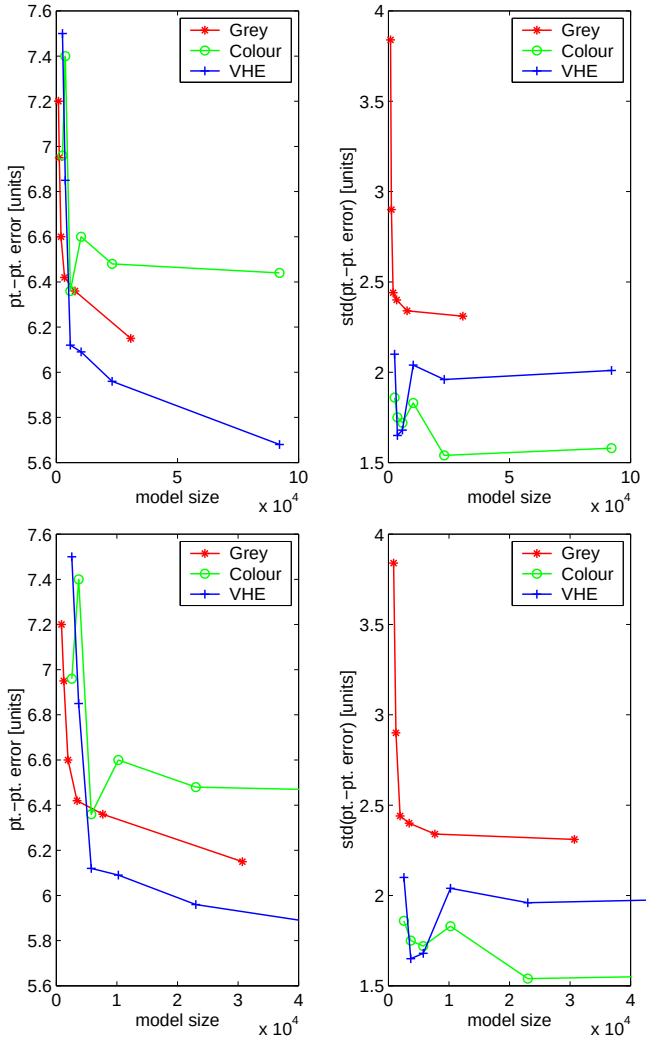

Fig. 6. AAM pt.-pt. accuracy measured as units vs. model size (left) and std. of units vs. model size in full (top) and zoomed (bottom) view.

In a typical benchmarking scheme pixel distances relates to a physical measure. In this experiment we define the physical measure a *unit* which is the width of a pixel at the input resolution, i.e. 640×480 pixels. The unit equivalent of figure 5 is shown in figure 6. From the zoom in figure 6 (bottom) it is seen that the VHE representation performs best for models larger than ∼5000 texture samples.

If the choice should only regard the resolution of grey-scale AAMs figure 6 (bottom) shows that the rate of improvement in unit accuracy is far smaller for models with more than ∼3000 texture samples.

Remarkably, the colour AAM had the lowest over-all unit accuracy and unclear trends in both unit mean and unit standard deviation plots.

VI. Implementation

All conducted experiments were based on an extended version of the AAM-API, which is an AAM implementation in C++ by one of the authors. A beta version of the AAM-API can be downloaded from `http://www.imm.dtu.dk/~aam/` This page also gives several examples on AAMs in other contexts.

VII. Discussion

Experiments have shown that a simple pre-processing of input images can increase segmentation accuracy on a limited set of facial images. The VHE representation outperforms conventional grey-scale AAMs and colour AAMs in cases with diffuse and partial directional illumination. The substantial – though not breathtaking dramatic – gain is obtained with negligible computational costs compared to colour AAMs but at the cost of a three times larger texture model compared to grey-scale AAMs.

Though stretched highly for display reasons, figure 2 indicates areas where the hue is ill-defined, e.g. at the eyes where the saturation is near zero. Further, in this discrete 24bit RGB setting, pixels near zero intensity also results in ill-defined hue angles. These areas could be learned from the training set and subsequently down-weighted. This would lead to better models.

Circumventing the need for two texture bands to represent the cyclic hue as shown may be too primitive. Colour ambiguity is introduced at all colour angels $\pm\theta$ measured from the point of collapse (in this case pure red at angle 0). However, in the presented case the loss in colour specificity is more than compensated by the gain from the over-all decrease in lighting sensitivity. This may not always be the case. For human faces though, hue is concentrated around the angle 0/360 [22]. For applications with limited and approximate unimodal distribution of hue – other than faces – the circular mean should be estimated (see e.g. [23]) to ensure a proper representation. For multimodal cases two texture bands should be used.

Concatenating all bands with a subsequent common linear normalisation as done in AAMs seem less optimal. In the VHE case the three subbands all have substantially different statistics suggesting that bands should be normalised separately using possibly non-linear means of nor-

malisation. This was earlier done on ultrasound images [10] with great success. In an initial stage this was applied to the edge band with limited success. The error increased due to the emphasis that was put on the noisy low to medium intensities, i.e. areas where the gradient are ill-defined. This could possibly be solved by a non-linear edge weighting scheme as suggested in [14] or by using a more elaborate regularisation prior to the gradient estimation, e.g. the anisotropic Perona-Malik diffusion scheme [24] or similar.

Finally, instead of patching the problem of non-Gaussian sources a more graceful solution would be to address the core of the problem. Namely, that PCA is based on assumptions of normally distributed variables. As such, Independent Component Analysis (ICA) [25] could prove to be a good replacement of the celebrated PCA.

VIII. Conclusions

Given the presence of colour information in a face segmentation task we have experienced the presented VHE representation to be an appealing alternative to model raw RGB intensities, in particular when dealing with change in lighting conditions. Using diffuse and partial directional lighting and at multiple scales, the VHE representation yielded higher accuracy than the conventional grey-scale and colour AAM. Only for very small models, grey-scale AAMs were more accurate, when measured relative to the size of the original input image. However, as absolute measures, the VHE performed best.

From the current experiments, the VHE representation should be preferred over the colour ditto. Further, compared to the grey-scale representation, the VHE should also be preferred if the required extra memory and computational power are available. For applications other than face segmentation, we have suggested modifications needed to utilise this intensity, hue and edge representation.

We have sought to promote the idea of modelling derived features combined with intensity information. Results showed that with subtle changes to a conventional grey-scale AAM framework and simple domain specific preprocessing a considerable increase in accuracy can be obtained. We anticipate that this also holds for other domains.

Acknowledgments

The following people are gratefully acknowledged for their help in this work. The face database was built by Michael Moesby Nordstrøm, Mads Larsen and Janusz Sierakowski. Dmitry Karasik made the multi-band extension of the AAM-API.

References

[1] S. Sclaroff and J. Isidoro, "Active blobs," *Proc. of the Int. Conf. on Comput. Vision*, pp. 1146–1153, 1998.

[2] J. Isidoro and S. Sclaroff, "Active voodoo dolls: a vision based input device for nonrigid control," in *Proc. Computer Animation '98*. 1998, pp. 137–143, IEEE Comput. Soc.

[3] T. Vetter, "Learning novel views to a single face image," *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pp. 22–27, 1996.

[4] M.J. Jones and T. Poggio, "Multidimensional morphable models: a framework for representing and matching object classes," *International Journal of Computer Vision*, vol. 29, no. 2, pp. 107–31, 1998.

[5] G.J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *ECCV'98. 5th European Conf. on Computer Vision. Proc.* 1998, vol. 2, pp. 581–95, Springer-Verlag.

[6] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," in *Proc. SPIE Medical Imaging 2001*. 2001, vol. 1, SPIE.

[7] S. Mitchell, B. Lelieveldt, R. Geest, H. Bosch, J. Reiber, and M. Sonka, "Time continuous segmentation of cardiac mr image sequences using active appearance motion models," in *Medical Imaging 2001: Image Processing, San Diego CA, SPIE*. 2001, vol. 1, pp. 249–256, SPIE.

[8] S.C. Mitchell, B.P.F. Lelieveldt, R.J. van der Geest, H.G. Bosch, J.H.C. Reiver, and M. Sonka, "Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac mr images," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 5, pp. 415–423, 2001.

[9] M. B. Stegmann, R. Fisker, and B. K. Ersbøll, "Extending and applying active appearance models for automated, high precision segmentation in different image modalities," in *Proc. 12th Scandinavian Conference on Image Analysis - SCIA 2001*, 2001, vol. 1, pp. 90–97.

[10] H.G. Bosch, S.C. Mitchell, B.P.F. Lelieveldt, F. Nijland, O. Kamp, M. Sonka, and J.H.C. Reiber, "Active appearance-motion models for endocardial contour detection in time sequences of echocardiograms," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 4322, no. 1, pp. 257–268, 2001.

[11] H. H. Thodberg, "Hands-on experience with active appearance models," in *Medical Imaging 2002: Image Processing, San Diego CA, SPIE*. 2002, SPIE.

[12] G.J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 1998, pp. 300–5, IEEE Comput. Soc.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. European Conf. on Computer Vision*. 1998, vol. 2, pp. 484–498, Springer.

[14] T. F. Cootes and C. J. Taylor, "On representing edge structure for model matching," in *Proc. IEEE Computer Vision and Pattern Recognition – CVPR*. 2001, vol. 1, pp. 1114–1119, IEEE.

[15] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. 1991 IEEE Com. Soc. Conf. on CVPR*. 1991, pp. 586–91, IEEE Com. Soc. Press.

[16] J. C. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, pp. 33–50, 1975.

[17] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–85, 1989.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[19] T. F. Cootes and C. J Taylor, *Statistical Models of Appearance for Computer Vision*, Tech. Report. Oct 2001, University of Manchester, http://www.isbe.man.ac.uk/~bim/, oct 2001.

[20] P. Switzer and A. A. Green, "Min/max autocorrelation factors for multivariate spatial statistics," Tech. Rep. 6, Stanford University, 1984, 10 pp.

[21] M. B. Stegmann, "Object tracking using active appearance models," in *Proc. 10th Danish Conference on Pattern Recognition and Image Analysis, Copenhagen, Denmark*. 2001, vol. 1, pp. 54–60, DIKU.

[22] G.R. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, , no. Q2, 1998.

[23] N. I. Fisher, Ed., *Statistical analysis of circular data*, Cambridge University Press, 1993.

[24] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," *Proceedings of the IEEE Computer Society Workshop on Computer Vision (Cat. No.87TH0210-5)*, pp. 16–22, 1987.

[25] P. Comon, "Independent component analysis – a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.