Identifying Individuals in Video by Combining 'Generative' and Discriminative Head Models

Mark Everingham and Andrew Zisserman
Department of Engineering Science, University of Oxford
{me,az}@robots.ox.ac.uk

Abstract

The objective of this work is automatic detection and identification of individuals in unconstrained consumer video, given a minimal number of labelled faces as training data. Whilst much work has been done on (mainly frontal) face detection and recognition, current methods are not sufficiently robust to deal with the wide variations in pose and appearance found in such video. These include variations in scale, illumination, expression, partial occlusion, motion blur, etc.

We describe two areas of innovation: the first is to capture the 3-D appearance of the entire head, rather than just the face region, so that visual features such as the hairline can be exploited. The second is to combine discriminative and 'generative' approaches for detection and recognition. Images rendered using the head model are used to train a discriminative tree-structured classifier giving efficient detection and pose estimates over a very wide pose range with three degrees of freedom. Subsequent verification of the identity is obtained using the head model in a 'generative' framework. We demonstrate excellent performance in detecting and identifying three characters and their poses in a TV situation comedy.

1. Introduction

The objective of this paper is to annotate video with the identities, location within the frame, and pose, of specific people. This requires both detection and recognition of the individuals. Our motivation for this is twofold: firstly, we want to annotate video material, such as situation comedies and feature films, with the principal characters as a first step towards producing a visual description of shots suitable for people with visual impairments, e.g. "character A looks at character B and moves towards him". Secondly, we want to add index keys to each frame/shot so that the video is searchable. This enables new functionality such as 'intelligent fast forwards', where the video can be chosen to play only shots containing a specific character; and character-based search, where shots containing a set of characters (or not containing certain characters) can easily be obtained.

The methods we are developing are applicable to any video material, including news footage and home videos, but here we present results on detecting characters in an episode of the BBC situation comedy 'Fawlty Towers'. Since some shots are close-ups or contain only face and upper body, we concentrate on detecting and recognizing the face rather than the whole body. The problem is thus essentially one of face detection and recognition. The task is a staggeringly difficult one. We must cope with large changes in scale: faces vary in size from 200 pixels to as little as 15 pixels (i.e. very low resolution), varying facial expression, partial occlusion, varying lighting, poor image quality, and motion blur. In addition, we must deal with detection and recognition of the face with *arbitrary* pose; in a typical episode the face of a principal character (Basil) appears frontal in one third of the frames, in profile in one third, and from behind in the other third. These imaging conditions are in distinct contrast to the classical domain of face recognition where factors including the camera placement, lighting, and facial expression, are typically controlled.

The approach we propose consists of three parts: (i) a 3-D model of an individual's face and head is built. This allows approximate images of the head to be rendered in novel views, giving extrapolation from the few training images provided. (ii) A tree-structured classifier is trained to detect the individual and estimate the pose over a very wide range of scale and pose. (iii) Initial estimates of pose are refined, and the identity verified using a generative approach and employing edge features and chamfer matching to give robustness to lighting and expression change.

1.1. Previous work

There has been much reported work on the problem of frontal face detection in still images [19, 23], with methods based on machine learning algorithms such as AdaBoost [23] giving reasonable accuracy with high computational efficiency. Such methods have also been applied to profiles [15, 19] and multiple views [13]. Li *et al.* [13] report experiments using a boosted pyramid of detectors trained to detect different out-of-plane rotations about the vertical axis (but not the horizontal axis), and exhaustive

(b) Face model

(a) Marked-up images (2 of 3 shown)　　(c) Head model
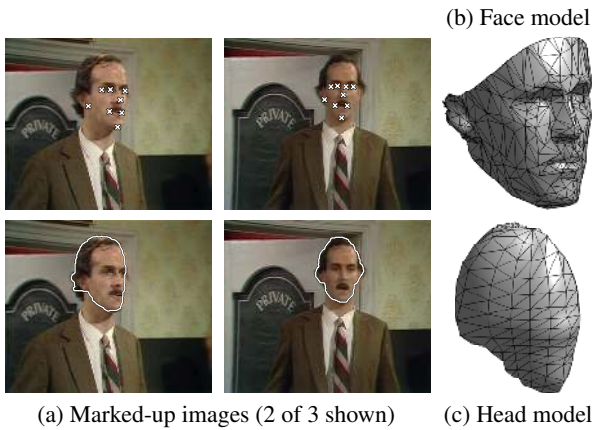
Figure 1. Modelling the face and head. A statistical deformable 3-D model is used to model the face, and a simple spherical deformable model is used to model the head. The pose and shape parameters are fitted to reference points and silhouettes in a few training images for a character to be identified.

Figure 2. Novel views of a character rendered using the 3-D head and face models in different poses and with different lighting. The images are rendered at the scale used by the detector.

search over in-plane rotation. Osadchy *et al.* train a convolutional neural network to simultaneously detect faces and estimate pose [16]. Detection of non-frontal views has not yet reached the same level of accuracy as the frontal case, principally because of the lack of rich visual features in non-frontal views of the face. In such views, visual features such as the hairline and occluding contour of the head are important for both detection and recognition.

Face recognition has similarly met with success in the case of frontal faces but performance lags behind for the case of variable pose. Recognition methods based on linear projections of the raw input image are typical, for example the 'eigenface' [22] or 'Fisherface' [1] methods. These require that the input images to be compared are registered to reasonably high precision, which cannot be achieved across pose variations. Methods proposed for dealing with variable pose can be categorized as view-based [5, 17], where the face is represented by a separate model for each of a finite set of poses, or 3-D model based [4, 18], where alignment of the 3-D model with the input image is used to factor out pose variation. Most work has used standardized face databases of high quality images, with little reported work on less constrained image data such as TV or movies. Several researchers have investigated clustering frontal faces in video [8, 11], leaving the task of naming the clusters to the user. Berg *et al.*[2] cluster frontal faces from news web sites and assign names to the clusters using the co-occurrence of a proper name in the accompanying web page text.

2. Approach

Instead of a paradigm of generic face detection followed by recognition, we build specific detectors for each person of interest to be found in the video. The aims of this approach are to be able to exploit visual features such as the outline of the head, which are distinctive for a particular person, making their use in generic face detectors problematic. From a few annotated training images we build a 3-D model of the person's face and head which can then be used to render novel images with different pose and lighting. Images rendered in this way are then used to train a discriminative tree-structured detector for the individual.

Because the detector operates on low resolution image patches and is trained to discriminate the head from background, rather than from other people, it lacks some specificity to an individual. We therefore apply a second refinement and verification stage which improves the pose estimate from the detector and gives a measure for assignment of identity to the detection.

2.1. Face and head model

The first stage of training in our approach consists of building a 3-D head model for each of the individuals to be identified. For the results reported here the video used is an episode of the 1975 BBC sitcom "Fawlty Towers" – because we are working with such archive footage, we have no opportunity to collect images of the characters in calibrated conditions, but must build the model from images available in the video.

Face model. The 3-D shape of the face is modelled using a statistical deformable model [4]. This model con-
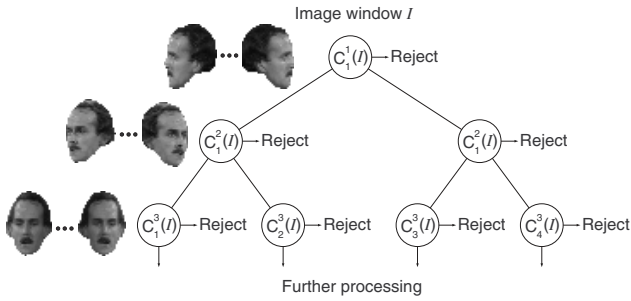
Figure 3. Detection and pose estimation using a tree-structured classifier. The classifier at each node detects a range of poses which is a subset of the parent.

a) Input   b) Edges

Figure 4. Edge features used as weak classifiers. The edge representation is shown at the pyramid level at which the character on the left (Basil) is detected. White pixels are non-edges and colors represent different orientations.

sists of a mean 3-D shape and a set of linear deformations of the shape derived by principal component analysis of laser-scanned faces and a corresponding linear model of appearance. Given hand-marked anchor-points, success has been obtained in fitting this model to single face images by stochastic gradient descent using an image-based error [4] or to multiple images using correspondence-based structure-from-motion methods [7]. We found these methods unsuccessful in this domain, primarily because of strong lighting effects and poor image quality. We therefore use a moderate amount of user supervision to fit the model.

Three images are used for model fitting, corresponding roughly to frontal, 3/4 and profile views. Two sources of information are used to fit the shape of the 3-D model: (i) reference points with known correspondence to points on the 3-D model are marked in each image, and (ii) the occluding contour of the face is marked in each image. The 3-D pose and shape parameters of the model are then estimated by minimizing the distance between the marked points and occluding contour and the projection of the corresponding model features in the image, subject to a Gaussian prior on the shape parameters [4]. We assume a weak-perspective camera, parameterizing the pose by 3-D rotation, 2-D translation and scale. The Levenberg-Marquardt with line-search optimization algorithm is used.

Figure 1a shows the marked points and silhouettes used to fit the model for one of the characters, and figure 1b the final fitted face model. Ninety principal components were used to represent the shape, and the triangle geometry of the model was decimated to 1,000 triangles (from the original 150,000 triangle model [4]) to reduce the computational expense of rendering.

Head model. For detection and recognition in non-frontal views, visual features including the hairline and shape of the head are important. These features cannot be captured by the face model since it extends vertically only to some part of the forehead, and in depth only to the ears. Equivalent

deformable models of the entire head are not common, and limited to the shape of the skull, because of the problems of capturing $360°$ range data, and the difficulty in modelling general hair. We assume that the hair can be considered reasonably rigid, which is valid for individuals with shorter hair, and that the shape of the head can be assumed reasonably smooth. These assumptions allow a generic 3-D shape model to be used without requiring a statistical model of the head derived from training data.

The head model is fitted using a shape-from-silhouette approach to recover the visual hull. We parameterize the shape of the head in a spherical coordinate system by the length of a ray from the origin to the surface of the model for a given azimuth and elevation. The pose of the model in each image has previously been estimated by fitting the face model, and the length of each ray is estimated by minimizing the distance between marked points on the boundary of the head in the image, and the silhouette cast by the model in the corresponding view. To regularize the solution, a prior penalizing high surface curvature is applied, using a discrete estimate of Mean curvature.

Figure 1c shows the final fitted head model. The model captures the overall shape of the head well, providing adequate accuracy to render the non-face region, but cannot capture features of the face, such as the nose, whose shape is under-constrained by the visual hull of the head.

Appearance model. To obtain more accurate rendering of the face region, including the correct prediction of self-occlusion, for example by the nose, the face and head models are combined. To render a novel view of the person, the image is first rendered using the head model, then the image rendered by the more-detailed face model is overlaid.

The appearance of the face is captured by back-projecting all three training images onto the 3-D model. New views of the model can then be rendered in different poses and with an approximation of different lighting – we assume a Lambertian lighting model, a single directed light

Figure 5. Example detections and initial pose estimates for three main characters in 'Fawlty Towers'. Note the wide range of scale and 3-D pose. Scale and pose estimates are approximate due in part to the granularity of scale and translation in the image pyramid and granularity of pose in the leaf nodes of the detector tree.

at infinity, and an ambient light source. Figure 2 shows examples of novel images of the model for one character rendered at the scale used by the detector (section 2.2).

In order to render examples of the face in different facial expressions, one might capture additional texture maps [9] or use a deformable 3-D model encompassing facial expression [3]. We take a different approach here. For detection, images of low resolution are used so that the effects due to changes in facial expression are small; for recognition, noting that the shape of the head is approximately invariant to facial expression, and that the overall location of features such as the eyes and mouth varies little with changes in expression, we instead use a measure for recognition which (a) exploits the shape of the head and hairline, and (b) has robustness to lighting and small deformations (section 2.3).

2.2. Detection

The 3-D model provides a reasonable representation of the appearance of the head over a wide range of poses. Previous work has applied the deformable face model directly to face recognition [4] by using the vector of shape and appearance parameters obtained by fitting as features for clas-

sification. This approach is infeasible for video applications because of the need to provide a good initialization for the appearance-based model fitting, typically by hand-marked points [4], due to the susceptibility to local minima. Several authors have proposed to solve the initialization problem by using detectors for local features of the face such as the eyes and corners of the mouth [6]. This approach is not easily applied to the low resolution images encountered in TV footage, in which the eye may be represented by just a few pixels, and for poses such as profile in which such local features are hard to detect.

The approach taken here is to build a discriminative detector for a particular individual using machine learning methods. Images rendered using the 3-D model provide the required training data, with backgrounds and negative examples taken as random patches from a large database of non-face images. This approach has several advantages: (a) features which are particular to an individual, for example the shape of the head, can be exploited to aid detection, and (b) the approach simultaneously provides detections and an estimate of the head pose.

Architecture. The detector is built using a tree-structured architecture exemplified in figure 3. The space of 3-D rotations is divided into successively smaller partitions using an octree structure (a binary tree is shown here for clarity). The root node corresponds to the full range of poses considered, being +/-90° azimuth and +/-30° elevation and in-plane rotation, and children represent a binary partition of each dimension of the parent range. The tree has 1,024 leaves corresponding to different poses. The additional pose dimensions of scale and translation are handled by scanning the detector over an image pyramid (a scale factor of $\sqrt{1.5}$ is used).

Each node consists of a classifier trained to detect images of the head in the corresponding range of poses. If the classifier responds to the input image, the children of the node are explored, else the entire branch of the tree is pruned. The tree architecture has two desirable consequences: (i) detection and pose estimation over a very wide range of poses is computationally efficient because of early pruning of the search; (ii) the accuracy of the detector is improved greatly by using a sequence of classifiers instead of a single classifier (each path through the tree can be considered a 'cascade' [23] with constrained structure). The tree architecture was inspired by the work of Stenger *et al.* on variable-pose hand detection [20], although in that case the training procedure and form of classifier differs significantly. It differs significantly from the pyramid architecture used by Li *et al.* [13] for multi-view face detection, in which nodes on a given level are arranged as a single cascade.

Classifiers and feature extraction. The classifier for each node is a linear combination of weak classifiers (re-

(a) Initial pose estimate from detector

(b) Refined pose estimate
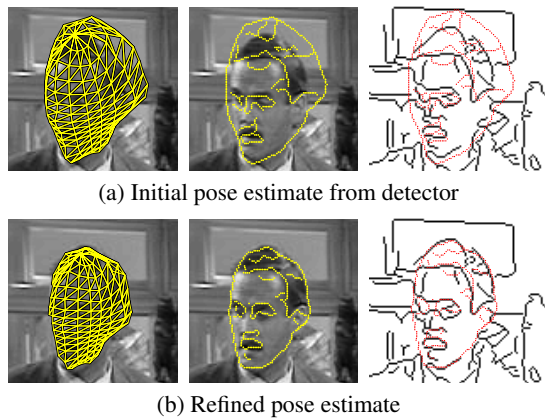
Figure 6. Pose refinement. The initial pose estimate (a) from the detector tree is refined by chamfer matching to give the final estimate (b). Columns show respectively the head model, rendered edges overlaid on the input image, and rendered edges overlaid on the edge image.

ferred to here as 'features' for clarity) trained using the AdaBoost algorithm [12]. The input to the classifier is a 45×45 image window; this is larger than the 20×20 windows typically used by frontal face detectors [23] but encloses the whole head rather than the face region alone (in the case of the 'Sybil' character, the hair occupies a large image area!). To minimize the execution time at both the training and testing stages, a relatively small and simple set of features is used [15] compared to the 'Haar-like' features often used in face detection [23]. The gradient of the image $\langle \delta_x, \delta_y \rangle$ is computed using symmetric finite differences, and the gradient magnitude $(\delta_x^2 + \delta_y^2)^{\frac{1}{2}}$ is thresholded; pixels with strong gradients are represented by the signed orientation of the gradient $\arctan(\delta_y, \delta_x)$ which is quantized into octants. Figure 4 shows an example of this input representation, in which each pixel takes on one of nine values: 'no edge' or orientation 1–8. This edge representation aims to capture the salient features of the image while diminishing the effects of lighting variation: it is invariant to additive changes in intensity (by removal of the DC component), and somewhat invariant to contrast variation (by thresholding), and gradual intensity gradients (because of the band-pass nature of the derivative).

Having computed the edge representation, the feature set $\{h_+(x, y, e), h_-(x, y, e)\}$ available to the boosting algorithm is simply the presence or absence of a particular edge type e at a given pixel (x, y) in the input window I:

$$h_+(x, y, e) = \begin{cases} +1 & : & I(x, y) = e \\ -1 & : & I(x, y) \neq e \end{cases} \quad (1)$$

where $h_-(x, y, e)$ is defined simply as $-h_+(x, y, e)$. This feature set is small (there are 18,225 possible features) and

very fast to compute; similar 'single pixel' features have successfully been applied to face detection by other authors [24].

The classifier $C_i^l(I)$ for each node of the tree is a linear combination of these single pixel features. If the output exceeds a threshold τ_i^l then the input window is passed to all eight children of a node, else the branch of the tree is pruned and produces no output. Classifiers at leaf nodes $C_i^n(I)$ for which the classifier output is above threshold output the sum of all classifiers along the path from the root to that node:

$$C_i^n(I) = \sum_{l=1}^{n} C_{\text{anc}(i,l)}^l(I) \quad (2)$$

where $\text{anc}(i, l)$ denotes the ancestor of node i at level l of the tree.

Figure 5 shows example detections for the three main characters in 'Fawlty Towers'. Note that the character is detected, and the pose approximately estimated, over wide ranges of scale (observe the image resolution in the original image) and pose, including profile views and rotation about three axes.

2.3. Pose refinement

The position, scale, and pose estimates produced by the detector tree are approximate for two reasons: (i) granularity of scale and translation due to the use of an image pyramid in the detector, and (ii) granularity of poses in the leaf nodes of the tree (around +/-5°). These pose estimates are refined by using the 3-D head model in a 'generative' mode.

Given an initial pose estimate, an image of the model in the corresponding pose is rendered and an edge detection algorithm applied. Edges detected in the rendered image are back-projected onto the 3-D model to obtain their 3-D coordinates, corresponding both to the occluding contour of the model and internal edges due to texture, for example the hairline. The matching error to edges in the input image is defined as a robust directed chamfer distance [10, 21, 18]:

$$d(U, V) = \frac{1}{|U|} \sum_{u_i \in U} \min \left(\min_{v_j \in V} ||u_i - v_j||, \tau \right) \quad (3)$$

where U is the set of model edge points and V is the set of input image edge points. Ambiguity of matches between edges is reduced by dividing each set of edges according to quantized edge orientation and only allowing matches of corresponding orientation. We allow matches with orientation error of around +/-22.5°; edges detected on the occluding contour of the model are allowed to match with edges of opposite orientation to account for the head appearing on either a light or dark background. The threshold τ makes the distance robust to some missing edges in the input image, for example spurious edges due to specularities on the

Figure 7. Example detections and identifications. Frames from a video sequence of 15 shots of 'Fawlty Towers' are shown with the head model of the identified character overlaid in the estimated pose. The three main characters are detected and identified over a wide range of scale and pose.

hair (see figure 6). Use of the chamfer distance rather than a pixel-wise grey-level/color measure as in other work [4] is advantageous for two reasons: (i) it is somewhat insensitive to lighting, and (ii) the 'slack' in the measure gives some robustness to changes in facial expression.

The initial pose estimate is refined by minimizing the chamfer distance using the LM-ICP algorithm [10] which uses the distance transform to make computing nearest edges efficient and Levenberg-Marquardt optimization. As the pose of the model is changed, three inaccuracies arise: (i) the edges corresponding to the occluding contour become inaccurate as this set is a function of pose, (ii) internal edges may be subject to self-occlusion, and (iii) the predicted orientation of the edges becomes inaccurate. We therefore run several passes of the algorithm, running to convergence then recomputing the model edge set.

Figure 6 shows an example of pose refinement; in this case the initial pose estimated by the detector (figure 6a) is fairly far from the correct one. After refinement by minimizing the chamfer distance, the boundary of the head model, hairline, and facial features match the input image closely (figure 6b).

2.4. Recognition

We noted in section 2.2 that the detector lacks some specificity to an individual due in part to the low resolution image patches used. As the final stage of the algorithm therefore, we verify the identity of each detection. For each detection, pose refinement is run using the 3D models for each of the characters of interest. The confidence that a detection is due to a particular character i is defined as:

$$
C(i) = \left[\frac{d(U_i, V)}{\min\left(\min_{j \neq i} d(U_j, V), \kappa\right)} \right]^{-1} \tag{4}
$$

where $d(U_i, V)$ is the chamfer distance (3) after pose refinement. Using the ratio between the distance to the character of interest and the nearest of the other characters gives a more informative score than the distance alone; use of this ratio has been suggested for matching invariant features [14]. The constant κ is introduced to reduce false positives on characters other than those modelled, and non-face detections.

We have obtained promising recognition results using this simple confidence measure. While we cannot claim that the measure would prove successful on a database of many people with very similar appearance, as might be encountered in classical face recognition, the TV domain considered here is somewhat different in that there are just a few central characters of interest.

3. Experimental Results

The proposed approach was evaluated on 4,400 keyframes (every tenth frame) of the episode 'A Touch of Class' of 'Fawlty Towers'. Models were built for three characters: Basil, Sybil and Manuel (see figure 5). Table 1 lists statistics of the test set. For each of the main characters, and all other people appearing in the video ('Other'), the number of faces are listed. For the main characters, the numbers of

(a) Basil (b) Sybil (c) Manuel
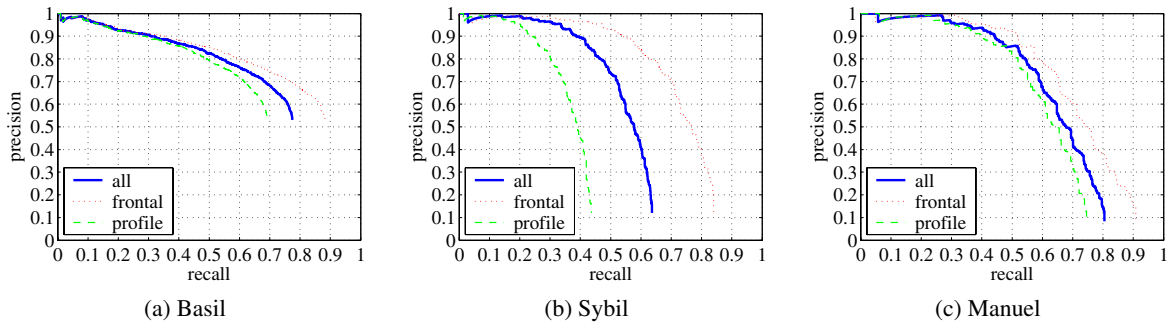
Figure 8. Precision/recall curves for the three main characters in 'Fawlty Towers'. A correct retrieval requires both detection and identification of the character. For each character, curves are shown for all views, 'frontal' only, and 'profile' only.

'frontal' and 'profile' views are given. The 'frontal' view was defined as both eyes being visible; there is therefore considerable pose variation even within the 'frontal' category. Frames for which a character is visible but not identifiable by their face i.e. facing away from the camera, are not included here.

Character	Frontal	Profile	Total
Basil	1,576	1,428	3,004
Sybil	415	416	831
Manuel	165	286	451
Other	–	–	3,023
Total			7309

Table 1. Statistics of faces in the test set. The number of faces for each character and view ('frontal' or 'profile') are listed.

For all characters, no more than around half of the images are in a frontal view, showing the importance of tackling detection and identification of non-frontal views in this domain. Around 40% of the faces in the video do not belong to one of the main characters, making this a very challenging test set.

Results on video. Figure 7 shows example output from our algorithms running on a sequence of 15 shots from this episode. The head model (excluding the face) is overlaid and the identity of each character labelled. For the majority of this sequence, two of the characters are in near-profile poses, challenging for current face detectors and recognition schemes. On average, the character's heads are around 60 pixels high in the image. Each character is detected and recognized correctly over wide ranges of scale, pose, and facial expressions.

The thresholds on the detector, chamfer distance (equation 3) and recognition measure (equation 4) were set empirically to give zero false positives on this sequence after detection and recognition (no non-faces or incorrect identifications). At this level of performance, the *detector* is still

generating a few false positives per image but these are successfully pruned from the output by using the recognition measure; this is in contrast to the isolated problem of face detection. Note that in the results reported here the algorithms are run independently on each frame – no 'tracking' is used. In the 180 key-frames of this sequence, with zero false positives or incorrect identifications, Basil is identified in 97%, Sybil in 81%, and Manuel in 98% of the frames in which they appear.

Quantitative results. Quantitative assessment of the proposed methods was conducted by treating the problem as one of retrieval, with the aim of retrieving all faces of a particular character. A correct retrieval requires both correct detection and identification of the character. Figure 8 shows precision/recall curves for each of the main characters tested. Recall is defined as the proportion of face images of a character retrieved, and precision is the proportion of the retrieved images which belong to the character of interest. For each character, three curves are shown: (i) retrieval of all faces covering all poses from profile to frontal; (ii) retrieval of all 'frontal' faces; (iii) retrieval of all 'profile' faces.

At a recall level of 50% the precision is around 80% for all characters. Results at higher levels of recall differ somewhat for each character, in part due to the varying number of images of each character (see table 1). For the characters Manuel and Sybil, who appear much less frequently than Basil, precision drops off above 50-60% recall, with confusion between the central characters and other people in the video increasing.

It is interesting to compare the precision/recall for retrieval of 'frontal' versus 'profile' views. As can be seen in figure 8, for the characters Basil and Manuel there is no clear preference for frontal views, with comparable curves for both subsets of the data. Subjectively, we judge that the head shape and hairline can provide strong cues to recognition in profile views. For the character Sybil however, there is a clear preference for frontal views; again subjectively,

we observed that in profile views of this character, the face region is poorly visible and the strong but unreliable texture in the hair, which occupies a large image area, caused problems both with fitting the model to the image and scoring matches.

4. Conclusions

We have proposed a method for detection and identification of individuals in video combining a 'generative' model with a discriminative detector, and utilizing an edge-based measure for pose refinement and recognition. Modelling the entire head, rather than just the face region allows our approach to exploit visual features such as the shape of the head and the hairline which are valuable cues for detection and recognition, particularly when the pose is far from frontal. Use of the chamfer distance for pose refinement and recognition gives some robustness to lighting, deformations caused by expression change, and inaccuracies in the model. Using these methods we have obtained very promising results on difficult unconstrained consumer video.

There are a number of ways in which this work may be extended. To deal with long hair it will be necessary to introduce more flexible hair models, which might be successfully done in the image domain rather than as a 3-D model. The similarity measure used for pose refinement and recognition could be extended with features beyond local edge orientation. Training a discriminative detector for each individual has enabled us to detect the face over a wide variation of pose which is still challenging for state-of-the-art face detectors, but this has associated computational expensive and would not scale well to hundreds of individuals, however the success of our approach suggests that training generic discriminative detectors from rendered images is a promising direction. In this work, we have not used temporal coherence either for model building, detection or identification, running independently on each frame. Exploiting the video domain in all stages should improve our results. Finally, incorporating other weak cues to identity such as clothing should be investigated, for example to cover frames where only the back of the head is visible.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE PAMI*, 19(7):711–720, 1997.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, pages 848–854, 2004.

[3] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Proc. EUROGRAPHICS 2003*, pages 641–650, 2003.

[4] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE PAMI*, 25:1063–1074, 2003.

[5] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *Proc. AFGR*, pages 227–232, 2000.

[6] D. Cristinacce and T. Cootes. A comparison of shape constrained facial feature detectors. In *Proc. AFGR*, pages 375–380, 2004.

[7] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *Proc. CVPR*, 2004.

[8] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content-based indexing of images and video using face detection and recognition methods. In *Proc. ICASSP*, 2001.

[9] M. Everingham and A. Zisserman. Automated person identification in video. In *Proc. CIVR*, pages 289–298, 2004.

[10] A. W. Fitzgibbon. Robust registration of 2D and 3D point sets. In *Proc. BMVC.*, pages 662–670, 2001.

[11] A. W. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. In *Proc. ECCV*, volume 3, pages 304–320. Springer-Verlag, 2002.

[12] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Proc. ICML*, pages 148–156, 1996.

[13] S. Z. Li, L. Zhu, Z. Q. Zhang, A. Blake, H. J. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *Proc. ECCV*, 2002.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*. Springer-Verlag, May 2004.

[16] R. Osadchy, M. Miller, and Y. Lecun. Synergenistic face detection and pose estimation with energy-based model. In *NIPS*, 2005.

[17] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. CVPR*, pages 84–91, 1994.

[18] S. Romdhani. *Face Image Analysis Using a Multiple Features Fitting Strategy*. PhD thesis, University of Basel, 2005.

[19] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. CVPR*, 2000.

[20] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Hand pose estimation using hierarchical detection. In *Proc. Intl. Workshop on HCI*, pages 105–116, 2004.

[21] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *Proc. CVPR*, 2003.

[22] M. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591, 1991.

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001.

[24] M.-H. Yang, D. Roth, and N. Ahuja. A SNoW based face detector. In *NIPS*, 1999.