Automated Visual Identification of Characters in Situation Comedies

Mark Everingham and Andrew Zisserman
Visual Geometry Group, Department of Engineering Science
University of Oxford
www.robots.ox.ac.uk/∼vgg/

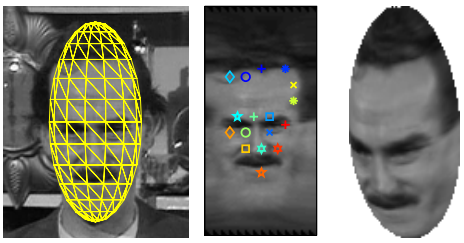
E-mail: {me | az}@robots.ox.ac.uk

Abstract

The objectives of the work described in this paper are simply stated: given examples of a particular person and an unlabelled video, we wish to find every instance of that person in the video and in others. This is an extremely difficult problem because of the many sources of variation in the person's appearance. We present a two stage approach. A 3-D ellipsoid approximation of the person's head is used to train a set of generative parts-based 'constellation' models which propose candidate detections in an image. The detected parts are then used to align the model, and the detections verified by global appearance. Novel aspects of the approach include the minimal supervision required and the generalization across a wide range of pose. We demonstrate results of detecting three characters in a TV situation comedy.

1. Introduction

The objective of the work presented here is to annotate video with the identities, location within the image, and pose, of specific individual people, requiring both detection and recognition of the individuals. Here we present results on detecting three characters in an episode of the BBC situation comedy 'Fawlty Towers'. As training data, we consider using just a single image of each person to be detected. 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 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), partial occlusion, varying lighting, poor image quality, and motion blur. In a typical episode the face of a principal character (Basil) appears in 1/3 of frames frontal, 1/3 in profile, and 1/3 from behind, so we have to deal with a much greater range of pose than is usual in face detection.



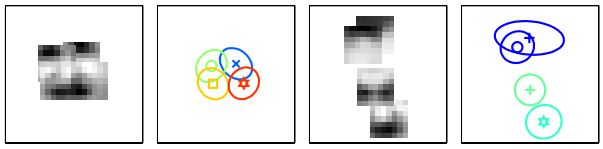
a) Ellipsoid b) Texture c) Novel view

Figure 1. Ellipsoid head model. Points on the texture mark the centres of selected parts on the ellipsoid

Our interest here is to locate characters in individual frames without the use of temporal information. This avoids the problems inherent with tracking, which are not our primary concern. We begin by synthesizing additional training data from a single training image by fitting a simple 3-D model to the person's head. Detection and recognition then proceed in two stages. Parts-based 'constellation' models [3, 6] of features on the head are used to propose candidate detections in an image. The 3-D model is aligned to the detected features, and global appearance used for verification.

1.1. Related work

Detection of frontal faces in images has reached some maturity in recent years, with several successful learning-based approaches [8, 14, 15]. Detection of faces with profile view [14] or arbitrary pose [8, 11] remains a challenging problem. There is a huge body of work on the related problem of face recognition [16]. Recognition of faces from high quality frontal images with stable illumination has met with some success using image-based 'eigenface' or 'Fisherface' approaches [1]. Recognition of faces with arbitrary pose and lighting remains challenging, with proposed methods based on alignment to a frontal view, multiple views [4], or 3D models [2]. There is some work on exploiting video for face recognition, using probabilistic tracking to improve



Frontal Facing bottom-left

Figure 2. Constellation models. Mean appearance is shown in mean position. Ellipses represent covariance of part positions. Covariance between parts is not shown.

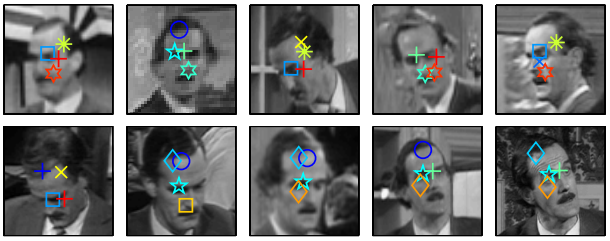


Figure 3. Constellations of parts detected across a wide range of poses using multiple aspects. Part symbols correspond to Figure 1b. A frontal face detector [14] fails to detect any of these faces.

robustness [10] or using the image trajectory as input for recognition [12]. Several researchers have also investigated automatic clustering of faces in video [5, 7].

2. Approach

2.1. Model definition

We build the model for each character from a single frontal image. The motivation for this is that such data could be collected automatically by frontal face detection and clustering [7]. An ellipsoid is fitted to the outline of the head, and a texture map for the model is obtained by back-projecting the training image onto the ellipsoid, allowing approximate novel views of the head to be synthesized. Figure 1 shows the ellipsoid model for the character Basil and the corresponding texture map. An example synthesized view is shown in Figure 1c.

The first task in identifying the individual is to detect the head and estimate it's pose. This might be achieved by search in the 6-D space of 3-D pose and rotation if a good initial pose estimate is available [2, 9], but here this requirement is avoided by using a view-based approach.

A multiple aspect version of the parts-based 'constellation' model [3, 6] is built from rendered images of the ellipsoid. An 'aspect' here is a subset of poses defined by covisibility of model features. The constellation approach allows the variation in appearance and shape (relative 2-D positions of parts), due to variation in pose or facial expressions, to be captured by modelling it probabilistically. For a single aspect v and a 2-D image position x, the probability of appearance A and shape S is factored by assuming each part appearance is independent of the shape and other part appearances, and the model is translation invariant:

$$p(A, S|v, x) = \prod_{i \in s_v} p(A_i|S_i, x)p(S|v) \qquad (1)$$

where s_v is the subset of parts which define a particular aspect v. The appearance densities $p(A_i)$ are shared across aspects but there is a distinct shape density $p(S)$ for each aspect.

To apply the constellation model, valid constellations of parts in an image are found by evaluating the joint log-likelihood of shape and appearance for each aspect around local maxima in the appearance log-likelihood. Search over scale is accomplished by using an image pyramid (1.2 scaling between levels is used here), and parts are located to sub-pixel precision by interpolation. The hypothesized constellations are ranked by log-likelihood and a subset of the most highly-ranked hypotheses retained for verification.

In the training stage, parts are automatically selected by choosing patches around 'interesting' points in the texture-map, and the viewing volume is automatically divided into aspects by analysing co-visibility of the parts. Full details are omitted here for lack of space. We consider views which are rotations up to +/-30^o about each 3-D axis, including combinations of both in-plane and out-of-plane rotations. Each part has size 7×7 pixels, where the distance between the eyes in the frontal image is just 7 pixels, and each aspect contains four wholly visible parts.

Figure 1b shows the position of the 17 parts selected for the Basil model. The probability distribution over part appearance is modelled using a PCA-based technique [13]. Image patches are represented as vectors and normalized to have zero mean and unit variance to obtain photometric invariance. These vectors are projected onto the first m principal components of the rendered set of parts, with m chosen to represent at least 80% of the variance in the data, and the distribution in this PCA space is modelled by a Gaussian with diagonal covariance. The distribution of the orthogonal components of the vectors is modelled as a spherical Gaussian. Inclusion of this residual term proves important.

The relative positions of the parts in an aspect is also modelled as Gaussian, treating the four 2-D part positions as a single 8-vector. All vectors are normalized such that the centroid lies at the origin, giving translation invariance, and the Gaussian has full covariance. Figure 2 shows the mean appearance, and shape distribution for two aspects of the Basil model, (a) roughly upright frontal views, the covariance shown capturing in-plane rotation, and (b) views of the character looking down to the (observer's) left.

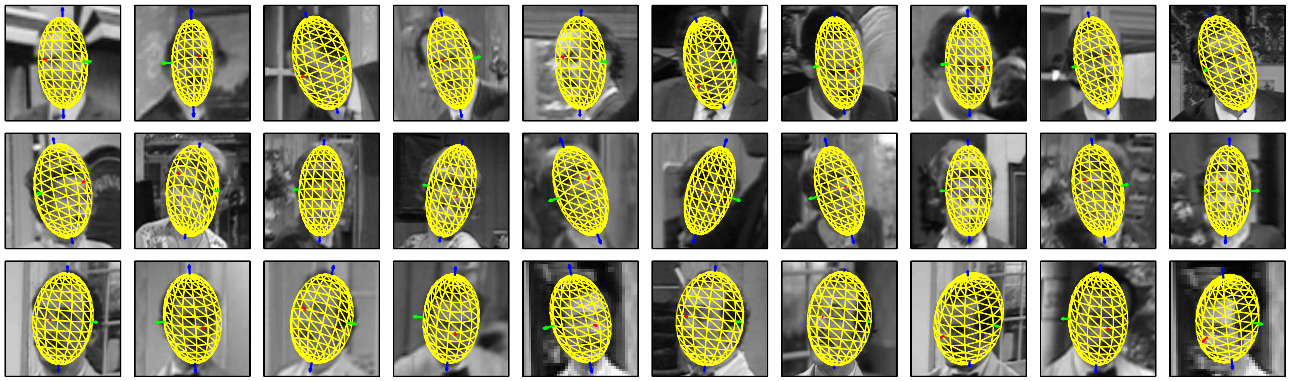


Figure 4. Detected characters with ellipsoid overlaid in estimated pose. The faces vary in size between 15 and 110 pixels.

2.2. Verification

Although the first stage detection model is built from data specific to a particular person, individuals cannot reliably be identified using the likelihood of the model alone. We therefore apply a second verification stage to hypotheses generated by the constellation model.

For each hypothesized constellation the corresponding pose of the ellipsoid model is estimated. Mathematically, assuming an affine camera and given the positions of four known parts in the image and the correspondence to the ellipsoid, the pose can be determined. Empirically we find it more robust to consider only those poses in which the parts are all wholly visible, which is determined by the detected aspect, and pick the pose which minimizes the sum of squared error to the detected part positions.

A gray-scale image of the ellipsoid model is then rendered in the estimated pose for comparison against the input image. Similarity measures such as normalized correlation between the images prove unsuccessful due to the small influence of the facial features, and instead we use a form of orientation correlation. Derivative operators are applied to the rendered gray-scale image T and visible points x having large gradient magnitude are selected. A measure of correlation between the edge orientations in the rendered and input images is then computed at the selected points:

$$c(I,T) = \frac{1}{n}\sum_{i=1}^{n}\frac{\nabla I(x_i)^\top \nabla T(x_i)}{|\nabla I(x_i)||\nabla T(x_i)|} \qquad (2)$$

where n is the number of selected points in the rendered image. This function falls off with the cosine of the angle between corresponding edges and is thus somewhat robust to outliers. The similarity measure can be thresholded to obtain a hard decision about the presence of an individual in an image, or sorted over frames or shots to give a ranking by likelihood that the person is present.

3. Experimental results

We tested our algorithm on 1,500 key-frames taken one per second from the episode 'A Touch of Class' of the BBC sitcom 'Fawlty Towers'. Detection of three of the main characters was evaluated: Basil, Sybil and Manuel. The task was to detect the frames containing each character, identify the image position of the face correctly (to within 0.3 of the inter-ocular distance) across pose variations exceeding +/-30^o about each 3-D axis, and correctly identify the character.

Figure 3 shows examples of Basil detected by the constellation model. This shows the strength of the approach, with the use of multiple constellations yielding successful detection over much wider variation in pose than attempted by current face detectors. A state-of-the-art frontal face detector [14] fails to detect any faces in these images. The detector also copes across a wide range of scale: faces shown are between 15 and 110 pixels wide, and lighting and motion blur in this video is particularly challenging due to 1975 video camera technology.

Figure 4 shows the ellipsoid model overlaid on the input images, with pose estimated from the detected parts of the constellation model. The first row corresponds to the parts shown in Figure 3, and the succeeding rows show images for the characters Sybil and Manuel. Qualitatively accurate pose estimates can be obtained across a wide range of views, and with image size as low as just 15 pixels.

Figure 5 shows receiver operating characteristic (ROC) curves for each of the three characters. Note that a correct identification requires both detection and recognition of the character, in contrast to face detection or recognition alone as in other work. We consider the results to be extremely encouraging given the difficulty of this task and the use of just a single training example per character. Identification of Sybil is particular good, with her being identified in over 80% of frames at a false positive rate less than 10%, and performance on the other two characters is also promising, and well above chance. It is noteworthy that 15–

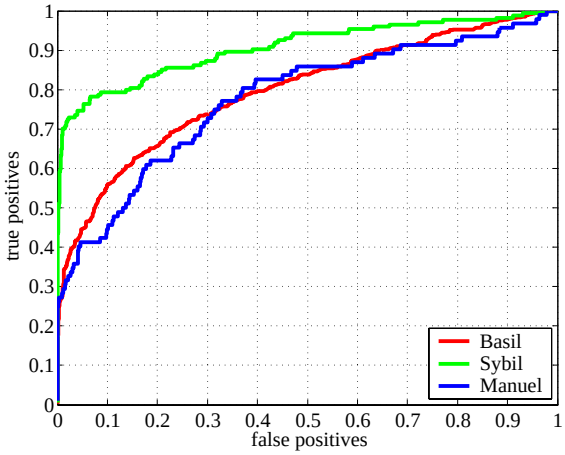


Figure 5. ROC curves for three characters in 1,500 key frames. Successful identification requires both correct detection and recognition.

Figure 6. The moustache problem. Two main characters and a secondary character '(Lord) Melbury' with similar appearance.

45% of key frames can be identified for each character with zero error, providing a useful starting point for 'bootstrapping' more robust models. Figure 6 offers some insight into the difficulty of identifying Basil and Manuel: their appearance, and that of a secondary character, is somewhat similar in a frontal pose at full resolution; in other frames where the face is as small as 15 pixels wide, and the pose far from frontal, distinguishing the characters is significantly difficult.

4. Discussion

We have presented methods for detecting and identifying characters in video across wide variations in pose and appearance. In future work we aim to expand the method to cope with full profile views and beyond, and improve robustness of identification with respect to variations in both pose and facial expression. We are investigating methods for unsupervised extension of the model presented here (see our web pages for a forthcoming publication on this subject), and more complete 3-D models. We currently consider the recognition of individuals only in those frames where we believe image evidence from the head supports it. In some shots of our test video, for example, Basil is present in the image, but facing completely away from the camera for the entire shot. To deal with such cases we need to investigate incorporation of less robust but complementary cues such as clothing (unstable since Sybil has a fondness for costume changes), and propagation of correct identifications by temporal reasoning.

Acknowledgements Funding for this work was provided by EC Project CogViSys.

References

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

[2] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illumination with a 3D morphable model. In *Proc. AFGR*, 2002.

[3] M. Burl, T. Leung, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. ECCV*, 1998.

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

[5] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content-Based Indexing of Images and Video Using Face Detection and Recognition Methods. In *Proc. IEEE ICASSP*, 2001.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, 2003.

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

[8] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. CVPR*, 2001.

[9] N. Krahnstoever and R. Sharma. Appearance management and cue fusion for 3D model-based tracking. In *Proc. CVPR*, June 2003.

[10] V. Krueger and S. Zhou. Examplar-based face recognition from video. In *Proc. ECCV*, 2002.

[11] 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.

[12] Y. Li, S. Gong, and H. Liddell. Video-based online face recognition using identity surfaces. In *Proc. IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, 2001.

[13] B. Moghaddam and A. Pentland. Probabilistic learning for object representation. In *Early Visual Learning*. Oxford University Press, 1996.

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

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

[16] W. Zhao, R. Challappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35, 2003.