Head Pose Estimation Using Multilinear Subspace Analysis for Robot Human Awareness

Tonislav Ivanov and Larry Matthies
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Dr. Pasadena, CA 91205
{tivanov lhm}@jpl.nasa.gov

M. Alex O. Vasilescu
Stony Brook University
Stony Brook, NY 11794
maov@cs.sunysb.edu

Abstract

Mobile robots, operating in unconstrained indoor and outdoor environments, would benefit in many ways from perception of the human awareness around them. Knowledge of people's head pose and gaze directions would enable the robot to deduce which people are aware of the its presence, and to predict future motions of the people for better path planning. To make such inferences, requires estimating head pose on facial images that are combination of multiple varying factors, such as identity, appearance, head pose, and illumination. By applying multilinear algebra, the algebra of higher-order tensors, we can separate these factors and estimate head pose regardless of subject's identity or image conditions. Furthermore, we can automatically handle uncertainty in the size of the face and its location. We demonstrate a pipeline of on-the-move detection of pedestrians with a robot stereo vision system, segmentation of the head, and head pose estimation in cluttered urban street scenes.

1. Introduction

Humans have the remarkable ability not only to estimate head pose of other people from far away, but also to use it to infer intent. Head pose alone provides a clue to the focus of attention of people, which is essential for human-computer interactions and intelligent autonomous robotic systems. Our goal is to enable mobile robots to observe head pose in cluttered, urban street environments to aid in predicting what the people might do next. However, the task of inferring the orientation of a human head from uncontrolled imagery is particularly challenging to a computer vision system. It requires invariance to lightning, identity, appearance, facial expression, and occluding objects. This task becomes even more difficult for distant subjects, where the head size is 32×32 pixels or less. Prior work in head pose classification at this low resolution is discussed in Section 2.

In this paper, we introduce a nonlinear, multi-factor model for head-pose recognition, that is based on the tensor algebra for appearance-based image analysis in [11]. Our model, described in Section 3, is invariant to identity and appearance of the people recognized, as well as to the scene illumination. It can estimate head pose of unfamiliar people in low-resolution images, where the head is as small as 15×15 pixels. Furthermore, this model handles scale uncertainty and localization by resizing itself and scanning over the image in a coarse-to-fine manner.

We tested the model on existing image database with cropped head chips in discrete horizontal (yaw) head poses, 10° apart. In out experiments, described in Section 4, we achieved pose classification accuracies ranging from 80% to 98%. We also applied the model on stereo image sequences acquired from a moving platform in busy urban streets. We used existing stereo vision-based pedestrian detection software to detect people, determine their range, approximately locate their heads, and do foreground/background separation prior to head pose estimation. In this very challenging scenario, discussed in Section 5, we obtained 57% correct classification of individual heads for left-center-right pose discrimination, with head sizes averaging 20×20 pixels. In the future, we plan to test in real-time and add tracking to improve results.

2. Prior Work

The vast majority of research in head pose estimation deals with relatively high resolution data. Many techniques, such as elastic bunch graphs and active appearance models, use facial features which are hard to detect in low-resolution images. Also, the stereo data on these distant images is too poor to build 3D models of the heads. Alternatively, tracking the heads can help in precise pose estimation, but requires initialization on stationary images. Single-

Ref.	Method	data	#poses	pose sep.	head size	classification rate	estimation error
[5]	auto-associative memory	Pointing	13	15°	23×30	50%	10°
[8]	neural nets images + edges	Pointing	13	15°	20×30	52%	9.3°
[10]	tensor model	Pointing auto-localize	13	15°	18×18 18×18	73% 55%	5° 12°
[4]	probabilistic model	PIE	9	22.5°	32×32 16×16 8×8	91% 85% 75%	
[4]	neural nets	PIE	9	22.5°	32×32 16×16 8×8	91% 87% 88%	
[9]	MLE + stereo	PIE indoor video	9 12	22.5° 30°	32×32 16×16 16×16	98% 90% 84%	
[7]	neural nets images + disparity	indoor video varying illum.	conti-nuous	±10°	24×32 24×32 varying	75% 60% 63%	7.5° 9.6° 9.7°
[14]	probabilistic model	video	4	45°	32×32		14.3°
[2]	decision trees	video	8	45°	10×10		37.9°

Table 1. Results on yaw angel head pose classification and estimation by various approaches on low-resolution images.

image pose detection can be done by using a set of appearance templates trained on specific poses; however, this approach is computationally expensive. The most promising techniques on single low-resolution images are appearance-based neural networks, support vector regression (SVR), and manifold embedding [6].

Some of the prior work on low-resolution images is summarized in Table 1. Classification accuracy and estimation error for yaw head pose are reported for the various methods. The most prevalent technique is neural networks (NNs). NNs have proven very successful, but [4] finds that NNs are very susceptible to head localization errors. To address this issue, [5] uses auto-associative memory, a type of NN, to perform pose estimation along with head detection for automatic alignment.

Most works report results on the Pointing Database (13 yaw poses, 15 ° apart) or the CMU PIE database (9 yaw poses, 22.5° apart). The images are usually scaled, alignment, and cropped prior to pose recognition. The best result on the PIE database of 98% accuracy is achieved by [9] using MLE and stereo on 32×32 images. The best result on the Pointing data of 73% is achieved by [10] using a tensor model at 18×18 resolution. The authors also report 55% recognition with automatic alignment. However, [10] as well as [8], only test on familiar subjects.

Results on images of non-stationary subjects taken from a stationary camera in an indoor environment are reported in [9]. After background subtraction and head region detection, the head pose is classified at 30° intervals with 84% accuracy. Similarly, a video of distant unfamiliar subjects with unrestricted movement is used in [7]. The authors perform color-base face detection as a means of segmentation, localization, and scaling. The yaw pose is predicted correctly in 63% of the heads within 10° of truth. Indoor and outdoor video data, with varying illumination and head size, was also used in [14] and [2]. Frames from a video of a person walking inside a lab are shown in [4], but no numerical results are given.

3. Model

In this paper, we employ a manifold technique called multilinear projection [13], which was similarly used in [10]. This method explicitly handles the embedding of a new sample from image into view space, where the pose is recognized. Our model can automatically handle localization and scale uncertainty, to which other methods, like NNs, are sensitive. It can be further combined with SVR to provide the dimensionality reduction on the image input. The formulation of the model is described next.

First, we organize the vectorized training images in a data tensor $\mathcal{D} \in \mathbb{R}^{I_{\mathrm{P}} \times I_{\mathrm{V}} \times I_{\mathrm{x}}}$, where I_{P} is the number of people, I_{V} is the number of views, and I_{x} is the number of pixels in an image. We apply multilinear analysis to decompose the data tensor as a N-mode product of orthog-

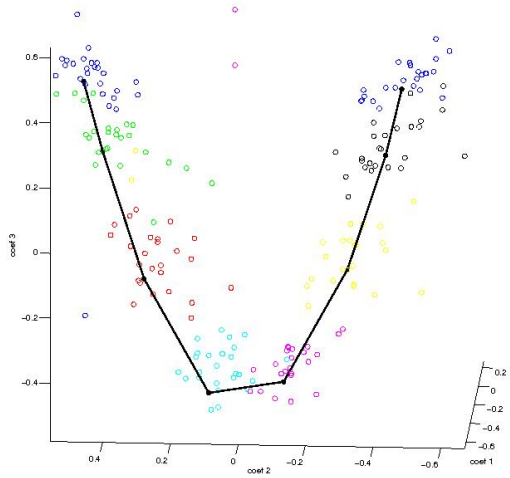

Figure 1. Viewpoint curve. All train images of the same view, map to a single point in view space (black dot). The black curve, connecting the views, follows the actual rotation of the head. The test images (small circles) project near the corresponding model view. Note that, for illustration purposes, only the first 3 dimensions of the 8-demensional view space are shown.

onal subspaces by a procedure called N-mode SVD [11]. The resulting representation separates the different modes of variation composing the facial image data: person identity, viewpoint, and appearance. We decompose $\mathcal{D}$ as:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{\text{P}} \times_2 \mathbf{U}_{\text{V}} \times_3 \mathbf{U}_{\text{x}} \quad (1)$$

$$= \mathcal{T} \times_1 \mathbf{U}_{\text{P}} \times_2 \mathbf{U}_{\text{V}} \quad (2)$$

where the core tensor $\mathcal{Z}$ governs the interaction between the modes, $\mathbf{U}_{\text{P}}$ spans the people space and contains row vector coefficients $\mathbf{u}_{\text{P}}^T$ for each person , $\mathbf{U}_{\text{V}}$ spans the viewpoint space and contains row vector coefficients $\mathbf{u}_{\text{V}}^T$ for each viewpoint, and $\mathbf{U}_{\text{x}}$ spans the image/pixel space and its columns are the conventional eigenfaces [13]. We can see the viewpoint space in Figure 1. All training images of the same view are projected to a single point in viewpoint space achieving zero intra-class scatter. The test images also project near the corresponding viewpoints.

The basis vectors $\mathcal{T}$ of our model, called tensorfaces (Figure 2), explicitly represent the variations of each mode across the image data and are expressed as:

$$\mathcal{T} = \mathcal{Z} \times_3 \mathbf{U}_{\text{x}} \quad (3)$$

$$= \mathcal{D} \times_1 \mathbf{U}_{\text{P}}^T \times_2 \mathbf{U}_{\text{V}}^T \quad (4)$$

where $\mathcal{T}$ is a third order tensor and is computed efficiently using equation (4).

A facial image $\mathbf{d}$ is expressed by a set of coefficient vectors, one for each mode: $\mathbf{c}_{\text{P}}$ is the person coefficient vector and $\mathbf{c}_{\text{V}}$ is the view coefficient vector. The image is represented as a multi-linear product of its coefficients and the tensorfaces:

$$\mathbf{d} = \mathcal{T} \times_1 \mathbf{c}_{\text{P}}^T \times_2 \mathbf{c}_{\text{V}}^T \quad (5)$$

For an unlabeled test image, we want to determine its mode coefficients in order to infer its mode labels, specifically the viewpoint. To do this, we construct a projection basis by taking the pseudo-inverse transpose of $\mathcal{T}$ flattened along the pixel mode:

$$\mathbf{P}_{[\text{x}]} = \mathbf{T}_{[\text{x}]}^{+T} \quad (6)$$

and then re-tensorize the projection basis $\mathcal{P}$ out of $\mathbf{P}_{[\text{x}]}$.

Next, we perform multilinear projection of the test image $\mathbf{d}$ from pixel space into the people and view mode spaces in order to simultaneously infer its coefficient vectors. To project the image, we multiply it with the projection tensor $\mathcal{P}$ to compute the response tensor $\mathcal{R}$ [13]:

$$\mathcal{R} = \mathcal{P} \times_3 \mathbf{d}^T \quad (7)$$

The response tensor $\mathcal{R}$ is a degenerate third order tensor of dimensionality $\Re^{1 \times I_p \times I_v}$ and of rank (1,1,1). To simultaneously infer the person and view representation, $\mathcal{R}$ is decomposed by N-mode SVD:

$$\mathcal{R} = \mathbf{c}_{\text{P}} \circ \mathbf{c}_{\text{V}} \quad (8)$$

Recognition is performed by nearest neighbor in viewpoint space using angle measure. The recognized view v is given by the smallest normalized scalar product between the inferred view coefficient vector $\mathbf{c}_{\text{V}}$ and a model's view coefficient vector $\mathbf{u}_v^T$, the vth row of $\mathbf{U}_{\text{V}}$:

$$\underset{v}{\text{argmin}} \left(\arccos \frac{\mathbf{u}_v^T \mathbf{c}_{\text{V}}}{||\mathbf{u}_v^T|| \, ||\mathbf{c}_{\text{V}}||} \right) \quad (9)$$

To represent the recognized by the model viewpoint, a reconstructed image $\mathbf{d}_r$ is computed according to equation (10), by multiplying the tensorfaces $\mathcal{T}$, the model's view coefficient vector $\mathbf{u}_v^T$ corresponding to the recognized view v, and the inferred person coefficient vector $\mathbf{c}_p^T$ of the test image. This reconstructed image is in one of the discrete viewpoints learned by the model, and resembles the likeness of the person in the test image to the best of the model's ability.

$$\mathbf{d}_r = \mathcal{T} \times_1 \mathbf{c}_p^T \times_2 \mathbf{u}_v^T \quad (10)$$

To measure how well the reconstructed image $\mathbf{d}_r$ matches the input image $\mathbf{d}$, we compute the residual vector $\mathbf{r}$ by taking the difference between the two images, but only inside the face region $\text{fx}(v)$, associated with the recognized viewpoint v :

$$\mathbf{r} = \mathbf{d}_r(\text{fx}(v)) - \mathbf{d}(\text{fx}(v)) \quad (11)$$

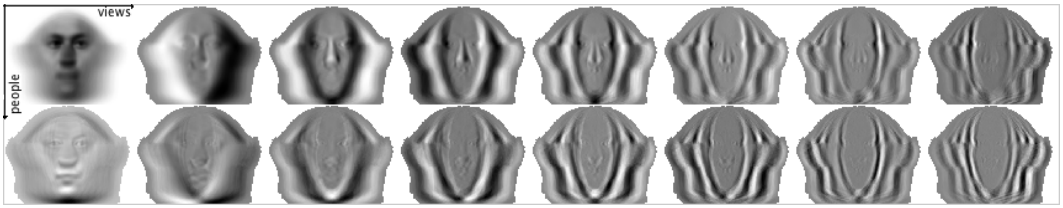

Figure 2. The TensorFaces of size 2×8×6344 for the reduced model with 2 people and 8 view coefficients. The view-axis captures viewpoint variation in the data set, while the people-axis captures people variation in the data set. The head pose of any image, in a view similar to the training views, can be represented as a tensor composition of these images. Note how the facial outline contributes to the view variation. The texture information of the face, which captures identity, is not important to represent a view.

From $\mathbf{r}$ we compute e, the root mean square re-projection error normalized by the number of face pixels $I_{\text{fx}(v)}$, as follows:

$$e = \frac{\sqrt{\mathbf{r}^T \mathbf{r}}}{I_{\text{fx}(v)}} \tag{12}$$

This error e is used to perform head localization and scale selection. It is largest when the input is a background image, and it is smallest when the input face is correctly aligned and scaled.

The tensor model separates each of the different modes underlying the formation of a facial image, enabling mode-specific dimensionality reduction. Since the head pose is independent of the person's identity, we reduce the people coefficients in our experiments using the prosedure in [12]. This makes the model more general and more compact.

4. Experiments

We conducted experiments with 80×107 pixel images of 75 people in 8 discrete viewpoints, 10° degrees apart (Figure 3), rendered from the Freiburg 3D morphable face database [3]. All faces were cropped, on white background, the same scale, and aligned at the pixel in the center between the eyes. The head pose was precisely measured. The results of the following experiments were averaged across 3-fold cross-validation. The training set contained 50 people in 8 views (400 training images), and the test set had 25 different people in the same 8 views (200 test images).

The model used in the experiments was derived by decomposition of the tensor of training images, as described in Section 3. We only trained on pixels that are part of the composite region occupied by meaningful face data in all training images. After the full model was trained, we drastically reduced the people coefficient representation from 50 to 2 to achieve viewpoint recognition invariant of the person's identity. Our model is represented by the 2×8×6344 tensorface $\mathcal{T}$ (Figure 2), 50×2 matrix $\mathbf{U}_{\text{P}}$, and 8×8 matrix $\mathbf{U}_{\text{V}}$.

In our base test, we achieved classification accuracy of 97.83% on aligned 80×107 test images. This result shows that the model's representation of viewpoint is indeed invariant of identity; thus, we can very well classify views of unknown subjects. In practice, however, such ideal condition, as a large high-resolution training set, and test images with perfect alignment, known scale, and uniform illumination, are rarely present. Next, we adapt the model to handle such situations.

Our model can learn from a small training set, unlike other models, like ANN and SVM, that require a large number of training images. Only 48 training images (6 people in 8 views) were sufficient for performance above 97% on the same test set as in the base test. Furthermore, the model was successfully trained on low-resolution images, as small as 15×21, and achieved recognition rate of more than 90% on same-resolution test images. Most importantly, the model has the ability to recognize images of different resolution by scaling to their size without retraining. The original model, trained on 400, 80×107 images, was able to scale down and recognize 90% of 200, 9×13 test images.

Our model can automatically localize a face of unknown scale in a uniform background image and simultaneously do pose recognition. To perform auto-localization, a search window is moved across the image, and the area within the window is projected onto the tensor basis. To accommodate for scale, windows of different sizes are scanned. The window with the lowest reprojection error, measured as described in Section 3, should contain the face correctly aligned, at the true scale, and in the correct face pose. To make the scan faster, a multi-resolution approach was implemented. The image was first scanned at lower resolution to determine the best location/scale match, and then locally, at higher resolution, to refine the result. We tested this procedure by taking each of the 200 original test images, scaling it randomly in size between 30×41 and 80×107 pixels, and then placing it randomly within a white background test image of size 240×321. First, we scanned that image at 1/3 resolution at all possible locations and scales, and then – at full resolution in a neighborhood of 10 pixels and 4 scale levels around the previously determined location and scale. We recognized the head pose in 79.5% of these images. This shows that our model can automatically handle

Figure 3. The 8 poses of a person from the Freiburg data, spaced from -35° to +35°, 10° apart.

both alignment and scale uncertainty.

When testing on database with illumination variation, the face pixels are normalized to a mean of zero and variance of one. This makes the dynamic range of the images the same, addressing lighting intensity variations. Also, our model can be extended to include a new illumination mode, so we can explicitly build invariance to illumination conditions. This, however, requires training on the same subjects in the same poses photographed in different lighting conditions. Due to significant image appearance differences, cause by the recording camera, between the Freiburg data and the data in our application, we limit training on the database on which we test.

We assume that in practice background subtraction can be provided by a head detection algorithm, such as skin color model or stereo segmentation.

5. Application

The data used to test real-world performance was gathered by a mobile robot driving on the sidewalk amongst people. A video sequence of 1000 frames was collected at 5Hz on 1024x768 imagery using stereo cameras with 60° FOV, which resulted in 15×15 heads at 10m. The ability to estimate head pose of people at this distance/resolution will allow the system to determine the people's future motion and their awareness of the robot.

We used existing stereo vision-based pedestrian detection software, described in [1], to detect people, determine their range, and do foreground/background separation prior to head pose estimation. The onboard pedestrian detection provides us with bounding boxes around the humans, from which we approximately locate their heads. From the average range to the people's heads, we infer approximate scale information. The stereo provides the head segmentation, but it is not perfect. It leaves in the image the person's torso, sometimes parts of the background, and sometimes cuts out parts of the face. Thus, the images of the heads which were extremely deformed were manually eliminated.

The model was trained on and limited to mostly upright frontal faces. We trained the model on images of 11 people in 3 pose (Figure 4) with resolution ranging from 13×13 to 63×63 pixels. All images were up-sampled to 80×107, and the heads were manually aligned and segmented. Missing side views were filled in by mirroring the opposing view. The face region of each image was normalized to a mean of 0 and a variance of 1 to reduce illumination variations.

Figure 4. The training images of 11 people in 3 poses from the robot video sequence.

The test set contained 256 images of 30 people, different from training, at resolution ranging from 15×15 to 61×61 (20×20 on average). The pose was manually labeled in three general categories: left, center, and right. The images belonging to a specific pose were not all in pure discrete directions, like in the Freiburg data, but rather a collection of unique views placed in pose categories. We applied the auto-scaling and localization procedure described in the previous section. We got 57% left-center-right head pose classification accuracy. Figure 5 illustrates the steps

Data	#poses	sep	resolution	class. rate
Freiburg	8	10°	80×80	98%
database			32×32	96%
			15×15	90%
auto align & scale			30×30 to 80×80	80%
Mobile robot Video data	3	45°	20×20	57%

Table 2. Out results on head pose classification.

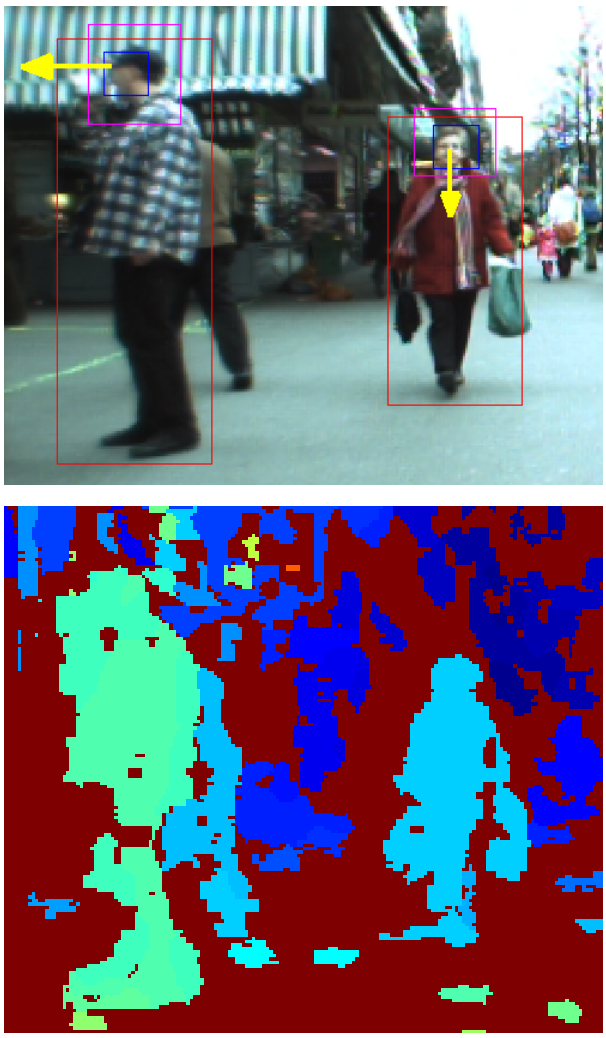

Figure 5. The top image is a 500×400 pixel cut out from the left rectified frame with red boxes around the detected pedestrians by the stereo system, magenta boxes around the head region on which we ran the pose detection, blue boxes indicating the automatically localized heads and their size, and yellow arrows pointing in the direction of the detected pose. The bottom image is the disparity calculated from stereo and used for background subtraction in the head region. The pose of the left-most person with head size 21×21 was detected as facing to the left. The pose of the other person with head size 17×17 was detected as facing forward.

of pedestrian detection, background subtraction, and head pose detection on a single video frame containing two of the test images.

Our results are summarized in Table 2. The results on the Freiburg data are quite good and serve as a validation of the proposed method. However, they cannot be compared directly with prior work which uses different data sets. Our video scenario is more challenging compared to prior work, because it was taken in a cluttered outdoor environment from a moving platform. Most prior work is on indoor videos of isolated subjects, which are taken from a stationary camera. Moreover, unlike prior work, our head detection procedure only approximately determined the head's position, size, and segmentation. Thus, we developed a model with the novel ability to automatically localize the face, handle uncertainty in scale, and tolerate some background. The successful pose recognition on difficult imagery, while performing these automated procedures, and given only several low-resolution training images, is quite remarkable.

6. Future Work

To improve the model, we plan to incorporate training on different illuminations. Also, we want to use multilinear ICA for decomposition of the data instead of the multilinear PCA currently used. This will help us achieve better results on image sets where the illumination is varying. For better viewpoint matching, we can experiment with different similarity measures and different metrics for reprojection error. This will help us perform better automatic localization and scaling, and allow us to simultaneous do head detection and pose classification without prior background subtraction.

For better situation awareness, we would like to extend the head pose estimation to include pan, tilt, and roll. In this case, the view space will look like a hemispherical surface instead of a parabola. Furthermore, to achieve more accurate pose estimation, we can expend our model to continuous poses by parameterizing the view space, for example by simply fitting a polynomial. Then, instead of finding the closest discrete view, we can find the closes point on the fitted curve and interpolate the pose.

To improve performance of the robotics system, we want to add tracking of the head, so we can use information from

previous frames as a prior on face location and pose. Such tracking can be handled with the tensor model itself.

7. Conclusion

Our model deals with the multiple variations inherent to image formation, such as identity, viewpoint, and illumination. We can determine head pose of unknown people in unconstrained environment. Out model is compact in size and efficient for computation, because all that is needed at testing time is projection of the face image onto the view subspace. We are able to classify head pose in low-resolution images, and automatically localize the face and determine its scale. We have successfully integrated automatic head pose detection into a stereo-based mobile robotic platform operating in a cluttered urban environment.

8. Acknowledgement

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

[1] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. H. Matthies. Results from a real-time stereo-based pedestrian detection system on a moving vehicle. In *Proc. of ICRA 09 Workshop on People Detection and Tracking*, 2009.

[2] B. Benfold and I. Reid. Colour invariant head pose classification in low resolution video. In *Proc. of BMVC*, 2008.

[3] V. Blanz and T. A. Vetter. Morphable model for synthesis of 3d faces. In *Proc. of SIGGRAPH 99*, pages 187–194, 1999.

[4] L. M. Brown and Y.-L. Tian. Comparative study of coarse head pose estimation. In *Workshop on Motion and Video Computing*, 2002.

[5] N. Gourier et al. Head pose estimation on low resolution images. In *CLEAR AFGR 2006 workshop*, 2006.

[6] E. Murphy-Chutorian and M. Manubhai. Head pose estimation in computer vision: A survey. *PAMI 2009*, 31(4):607–626, 2009.

[7] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proc. of AFGR*, 2004.

[8] R. Stiefelhagen. Estimating head pose with neural networks – results on the Pointing04 ICPR workshop evaluation data. In *Pointing04 ICPR workshop*, 2004.

[9] Y.-L. Tian, L. Brown, and J. Connell. Absolute head pose estimation from overhead wide-angle camera. In *Workshop on Analysis and Modeling of Face and Gestures*, pages 92–99, 2003.

[10] J. Tu and T. Huang. Locating nosetips and estimating head pose in images by tensorposes. In *Proc. of ICIP 2007*, pages 513–516, 2007.

[11] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. of ECCV 2002*, pages 447–460, 2002.

[12] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. Of CVPR 2003*, pages 93–99, 2003.

[13] M. A. O. Vasilescu and D. Terzopoulos. Multilinear projection for appearance-based recognition in the tensor framework. In *Proc. Of ICCV 2007*, pages 1–8, 2007.

[14] Y. Wu and K. Toyama. Wide-range, person- and illumination-insensitive head orientation estimation. In *Proc. of AFGR*, 2000.