Colour Invariant Head Pose Classification in Low Resolution Video

Ben Benfold and Ian Reid
Department of Engineering Science
University of Oxford, OX1 3PJ, Oxford, UK
{bbenfold,ian}@robots.ox.ac.uk

Abstract

This paper presents an algorithm for the classification of head pose in low resolution video. Invariance to skin, hair and background colours is achieved by classifying using an ensemble of randomised ferns which have been trained on labelled images. The ferns are used to simultaneously classify the head pose and to identify the most likely hypothesis for the mapping between colours and labels. Results from video sequences demonstrate that an improved posterior estimation using learnt colour distributions reduces classification error and provides accurate pose information in images where the head occupies as little as 10 pixels square.

1 Introduction

In systems which automatically monitor surveillance video, knowledge of head pose provides an important cue for higher level behavioural analysis. The focus of an individual's attention often indicates their desired destination whereas mutual attention between people indicates familiarity, and any single object or person receiving attention from a large number of people is likely to be worthy of further investigation. In systems controlling dynamic cameras, a pose estimation from a low resolution head image can be used to determine whether or not a close-up from a dynamic camera would provide a face image that is suitable for identification.

Surveillance cameras tend to have a fairly wide field of view, making the region of the video that is occupied by a person's head fairly small. The low resolution of the head image prevents the application of techniques which require detail such as those which track feature points or detect facial features [6, 4]. The majority of research into head pose measurement in low resolution video involves the use of labelled training examples which are used to train various types of classifiers such as neural networks [11, 2, 13], support vector machines [14] or nearest neighbour and tree based classifiers [10, 7, 1]. Other approaches model the head as an ellipsoid and either learn a texture from training data [15] or fit a reprojected head image to find a relative rotation [9].

For a head pose classifier to be effective in real-world situations it must be able to cope with different skin and hair colours as well as wide variations in lighting direction, intensity and colour. Most existing classifiers are susceptible to these variations and require examples with different combinations of lighting conditions and skin/hair colour variations in order to make an accurate classification.

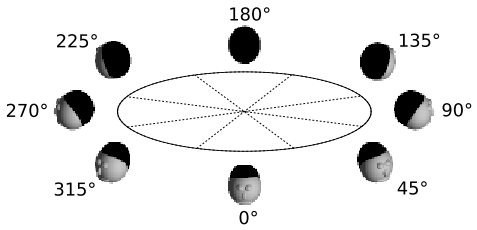


Figure 1: The eight head pose classes into which all head images were classified based on the pan angle of the head. Pan was measured about the y-axis of the image and for heads with rotations about other axes pan was considered the first applied.

This work bears closest comparison to that of Robertson [10], who tried to classify gaze direction into one of eight different classes. Although that work showed some promising results, the training and classification results were critically dependent on a good model of skin colour, and yielded unreliable results which needed to be smoothed with a prior based on the direction of motion of the person. In contrast, the approach presented in this paper effectively learns a model of the skin colour for each new video, making it largely invariant to lighting and person dependent characteristics of the video, and as such does not use any other cues.

The method described classifies the head images into the same eight direction classes as Robertson (figure 1) using a randomised fern classifier, with invariance to the skin and hair colour of individuals being achieved through the use of segmented head images. Section 2 demonstrates a method by which randomised ferns can be used to simultaneously classify the head direction and label each segment as either hair, skin or background. It is then shown that the classification rate can be improved by automatically learning the hair, skin and background colours without initialisation or prior assumptions. Section 3 describes how the output of the classifier can be filtered through the use of a Hidden Markov Model (HMM) to remove random errors. The results of experiments using a series of video sequences are analysed in section 4 and concluded in section 5.

2 Randomised Ferns

2.1 Training

Randomised ferns [8] are a type of decision tree where the decisions for branches at equal depths are all the same, which allows a more efficient implementation than a standard decision tree. The ferns are first built by randomly selecting branch decisions and are then trained using example images for which the correct class is already known. The training examples are passed down the fern until they reach a leaf node corresponding to the outcome of all of the branches, where they are added to a histogram bin corresponding to their class. The normalised histograms at the leaf nodes represent the relative probabilities that an image will belong to each class given that it has arrived at the leaf.

In our case an ensemble of randomised ferns was trained using approximately 1000 images which had been manually assigned to one of eight classes based on the head orientation. The training images were resized to ten pixels square and converted to a normalised YUV colourspace to reduce the distance between colours differing only in

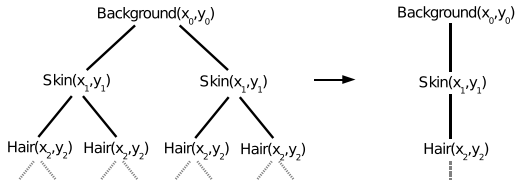


Figure 2: Ferns are decision trees with the same decisions at each level regardless of previous branch outcomes, allowing an efficient implementation where the same tests are applied to all images. Predicates which test image locations for the presence of labels were used as the decisions for the head classifier.

luminance. The images were then divided into six colour segments using k-means clustering. Six segments were used to ensure that the hair and skin were separated in cases where lighting conditions cause their colour distributions to be multimodal. Each segment was manually assigned one of three labels: *hair*, *skin* or *background* and it is these labels upon which the branch decisions for the ferns were based. The branches test for the truth of a predicate applied to a single randomly chosen pixel (figure 2). The experiments detailed in section 4 were performed using twenty ferns with sixteen branches each.

When the ferns are used to classify images from video, the correct hypothesis for the labelling of segments is not known in advance so the ferns must be used to test the probability that a labelling hypothesis is correct or incorrect for a particular segmented head image. The total number of samples in the histogram at each leaf node represents the probability that any correctly labelled image will reach that leaf. Unfortunately, incorrectly labelled images are more likely to reach some leaf nodes than others due to structural information being lost when multiple segments are given the same label. An extreme example would be a hypothesis which labels all segments as hair; under this hypothesis all images would reach the same leaf. To prevent bias towards these simple hypotheses, it is important to take into account the probability of a leaf being reached by both correctly and incorrectly labelled images.

To represent the probabilities of misclassification, a second set of histograms were populated at each leaf using the training images under every possible incorrect labelling hypothesis. This second set of histograms represents the probability that an image reaching the leaf would be allocated to each class given that it has been labelled with an incorrect hypothesis.

2.2 Classification

To recover the correct labelling hypothesis for a segmented head image I, different hypotheses must be tested using the randomised ferns. Let H and C be the sets of all hypotheses and all classes respectively, with $P(h,c), h \in H, c \in C$ being the joint probability that c and h are both correct for I. If d_h^f represents the outcomes of the branches in fern f from a total of n when I is labelled using hypothesis h then the joint probability can be calculated by expanding Bayes' formula with the assumption that $P(h)$ and $P(c)$ are independent:

$$P(h,c|d_h^f) = \frac{P(d_h^f|h,c)P(c)P(h)}{\sum_{c_i \in C} P(d_h^f|h,c_i)P(c_i)P(h) + P(d_h^f|\overline{h},c_i)P(c_i)P(\overline{h})} \quad (1)$$

The joint distribution can be marginalised and the mean taken over all n ferns to give estimates of the probability that each hypothesis and class is correct for the image:

$$P(h|d_h^1 \dots d_h^n) = \frac{1}{n} \sum_{1 \leq f \leq n} \sum_{c_i \in C} P(h, c_i | d_h^f) \tag{2}$$

$$P(c|d_h^1 \dots d_h^n) = \frac{1}{n} \sum_{1 \leq f \leq n} \sum_{h_i \in H} P(h_i, c | d_h^f) \tag{3}$$

The values of the conditional probabilities $P(d_h^f|h,c_i)$ and $P(d_h^f|\overline{h},c_i)$ are represented in the histograms from the leaf corresponding to d_h^f. It is standard practise to consider the output of each randomised fern to be independent and multiply the output probabilities to give the combined estimation. In our case a single labelled image contains a maximum of $\log_2 3^{100}$ (≈ 158) bits of information whereas the ensemble of ferns makes a total of 320 binary tests, so there is a large amount of mutual information between the estimations from the ferns. This mutual information causes the product of the estimations from the ferns to be severely biased, so instead the estimations were combined by taking the mean over all ferns in the same way that one would for a forest of randomised trees.

2.3 Learning Colour Distributions

The segment structure alone provides enough information to classify the head image successfully, however the classification accuracy can be improved by learning the colours represented by the labelled image segments and using them to calculate priors for a second search of the ferns. Although no assumptions are made about the hair and skin colour distributions of the surveillance subject, it is quite safe to assume that they will not change significantly while they are being observed.

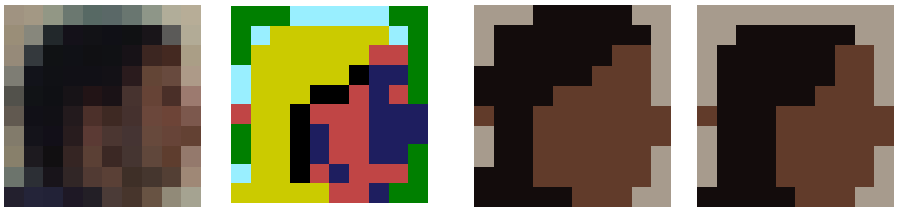
The colour distributions for each label are learnt by taking the hypothesis which maximises the probability in equation 2:

$$h_{max} = \underset{h}{\operatorname{argmax}} P(h|d_h^1 \dots d_h^n) \tag{4}$$

In every frame, the pixel colours from each of the segments in the head image are added to the histograms to which they are labelled by h_{max}. The resulting colour histograms $Y = \{Y_{hair}, Y_{skin}, Y_{background}\}$ that are built up over all frames up to and including the current one allow the hypotheses for the current frame to be weighted by comparing the colour distributions of each segment $s \in S$ to each of the colour histograms. The similarity of the histograms is estimated using the Bhattacharya coefficient and the overall probability that a hypothesis is correct is proportional to the product of the coefficients for each segment:

$$P(h|Y) \propto \prod_{s \in S} B(s, Y_{h(s)}) \tag{5}$$

In the above equation, the notation $Y_{h(s)}$ is used to represent the colour histogram corresponding to the label that h provides for segment s and $B(s, Y_{h(s)})$ represents the Bhattacharya coefficient calculated between the two histograms. Equation 1 can be adjusted to provide an improved estimation by taking into account the colour histograms Y:



(a) Scaled head image (b) The six segments (c) The most likely hypothesis (d) The most likely hypothesis using colour

Figure 3: Example head image segmentation and the most likely hypotheses both with and without the use of colour information. The hypothesis in image (c) is the most likely based on the structure of the segments alone and the more accurate hypothesis in (d) resulted from classification with the hypotheses weighted using the learnt colour distributions.

$$P(h,c|d_h^f,Y) = \frac{P(d_h^f|h,c)P(c)P(h|Y)}{\sum_{c_i \in C} P(d_h^f|h,c_i)P(c_i)P(h|Y) + P(d_h^f|\overline{h},c_i)P(c_i)P(\overline{h}|Y)} \tag{6}$$

Equations 2 and 3 can be updated similarly to give improved estimates of the class and hypothesis probabilities.

The use of colour as a temporal constraint allows ambiguities to be resolved in situations where structure alone does not provide enough information. An example of this is shown in figure 3 in which the most likely hypothesis for the segmented image gives the entirely plausible labelling shown in figure 3(c) which causes the image to be misclassified. The priors obtained using the colour histograms allow the calculation of an improved posterior hypothesis, giving the more accurate labelling shown in figure 3(d) which results in correct classification.

3 Hidden Markov Model

The pose of a head at any time provides information relating to the pose shortly after due to the physical constraints of human head motion. This information can be exploited through the use of a Hidden Markov Model (HMM) to filter the head pose estimations. If each of the head pose angles are modelled as individual states, the probability of transitioning from state i to state j can be estimated as a Gaussian function of the smallest positive angle a_{ij} between them:

$$T_{ij} \propto e^{-\frac{a_{ij}^2}{2\sigma^2}} \tag{7}$$

The observations from the frame at time t can be written as a diagonal matrix O_t where the ith diagonal element represents the probability that the head is in state i given the output of the classifier at time t. If s_t is a vector representing the probability that the head is in each state at time t, a filtered estimation of the state probabilities at time $t+1$ can be calculated from the observation and state transition matrices:

$$s_{t+1} \propto O_{t+1} T s_t \tag{8}$$

This model is however based on the assumption that the observations are all made at equal intervals which is often insufficient. Cameras operate at different frame rates and one cannot guarantee that every frame will be processed in a real-time system, so to cope with unequal observation intervals equation 8 was modified to allow an arbitrary observation interval Δt:

$$s_{t+\Delta t} \propto O_{t+\Delta t} T^{\Delta t} s_t \tag{9}$$

In this case the Gaussian distribution of T allows $T^{\Delta t}$ to be calculated simply by changing the standard deviation in equation 7 to be a function of the time interval. If k is a constant representing the estimated standard deviation of the angular velocity, σ can be calculated using:

$$\sigma = k\sqrt{\Delta t} \tag{10}$$

The use of this model allows filtering to be performed when frames are missed and ensures that the constant k need not be adjusted for cameras with different frame rates.

4 Results

The classifier was used to estimate the head pose in a total of 9260 head images from a set of videos including sequences from the Hermes and Terrascope datasets [3, 5]. The test videos cover a wide variety of different lighting conditions and include images from sixteen actors with different hair and skin colours. Sample frames from the videos are shown in figure 4. Ground truth for each head image was hand labelled using high resolution frames, so the angular errors calculated using it are likely to be overestimates.

The use of only eight classes results in the pan angle estimations being severely quantised. A more accurate pan angle estimation was found by taking the rotation of the class probabilities at which the convolution with a Gaussian gave the largest response.

The absolute angular errors (figure 5) demonstrate that head pose can be recovered to a fairly high degree of accuracy despite the low resolution. The mean absolute angular errors for the classifier with and without the colour constraints were 38.4 and 43.5 degrees respectively, demonstrating a clear improvement. With the addition of the HMM filtering, the mean error dropped to 37.9 degrees. Whilst the HMM does not significantly reduce the mean error, it does filter the pose estimations over time which makes small head motions more easily identifiable (figure 6).

The main limitation of the approach is the requirement that the head location must be known in each frame. In most cases the head could be found automatically, however any inaccuracies increase the error in the head post estimation (figure 7). Fortunately the randomised ferns can be used to help localise the head region in situations where the location is inaccurate . The sum of all hypothesis probabilities provides a measure of the likeliness that an image is a head, so by evaluating the hypotheses with the head region translated by small amounts the optimal location can be found. Figure 8 shows the results from an experiment where the hypotheses were evaluated with the head region shifted to six different locations. Although the ferns do not always find the correct location, they do provide a significant amount of information which could be used to aid localisation.

People are capable of identifying frontal face views in very low resolution images [12], however non-frontal views tend to be less distinctive and in many cases people find the identification of the head pose challenging. To provide a comparison with human

Figure 4: Sample frames from three video sequences with estimated head pose angles annotated. Although these frames are shown in high resolution, the head images were scaled to ten pixels square before classification. In the bottom sequence, the actor on the right has only hair visible for the majority of the video which prevents the colour distributions from being learnt correctly and results in large errors.

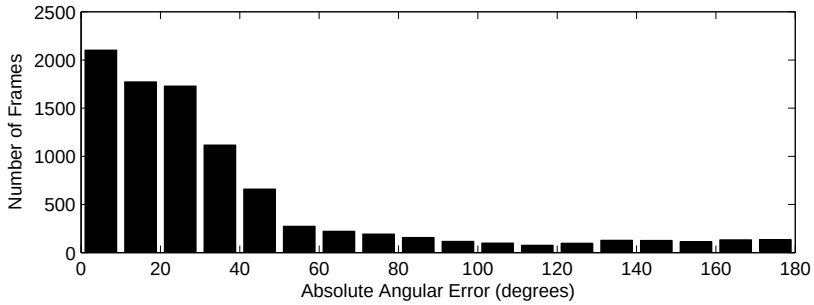
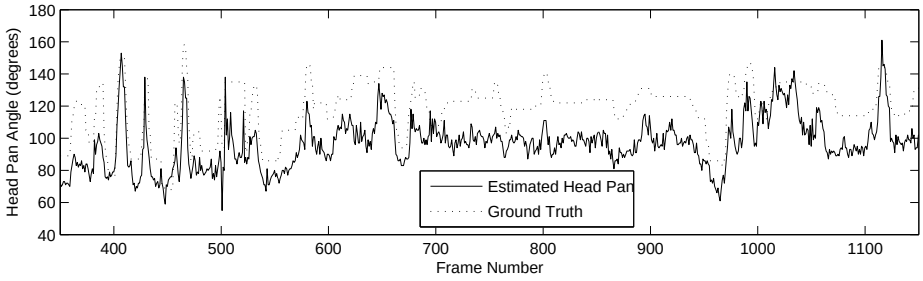
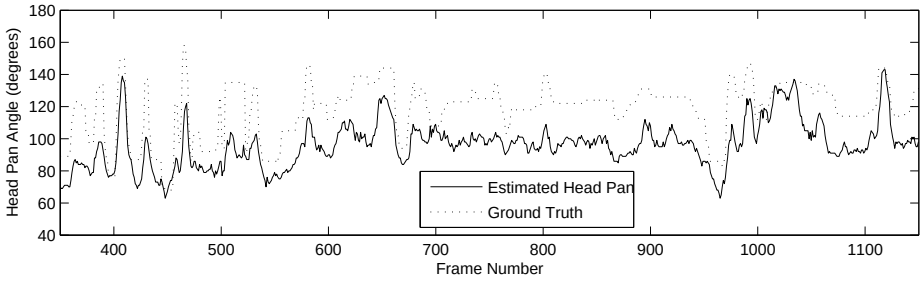


Figure 5: Angular error histogram from tests using 9260 head images, the mean being 37.9 degrees. Small angular errors are common due to the similarity of adjacent poses

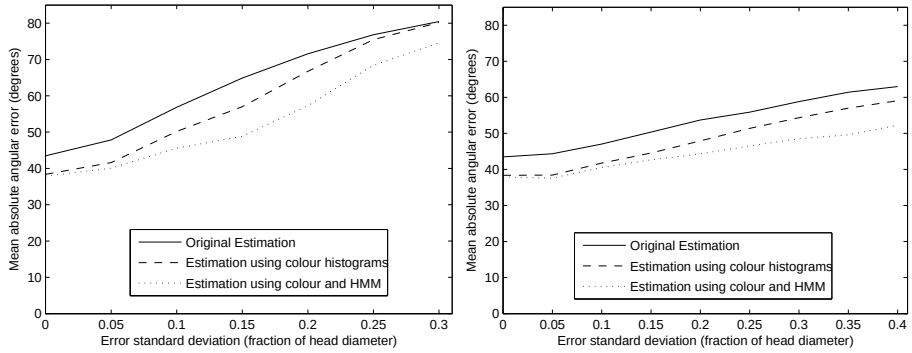


(a) Head Pose estimation without filtering



(b) Head Pose estimation filtered using a HMM

Figure 6: Pose estimation graphs for a single head compared with ground truth both with and without filtering through the use of a Hidden Markov Model. The filtering reduces the amount of noise and makes small changes in the head pose more visible.



(a) Head pose error resulting from the head region being incorrectly positioned

(b) Head pose error resulting from the head region being too large or too small

Figure 7: Graphs showing the effects of random translational and scale errors on the accuracy of the head pose estimation. Errors were introduced with a Gaussian distribution and resulted in the head region being translated or scaled along both axes. The classifier performs well when the size of the head region is incorrect but performance rapidly diminishes as translational errors are introduced.

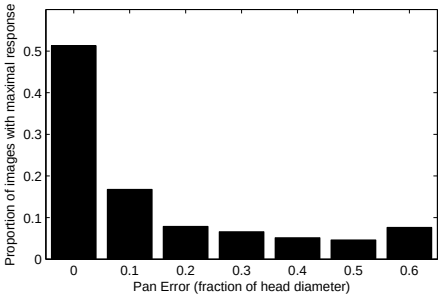


Figure 8: The fraction of head images for which the randomised ferns estimated the greatest probability at each pan location.

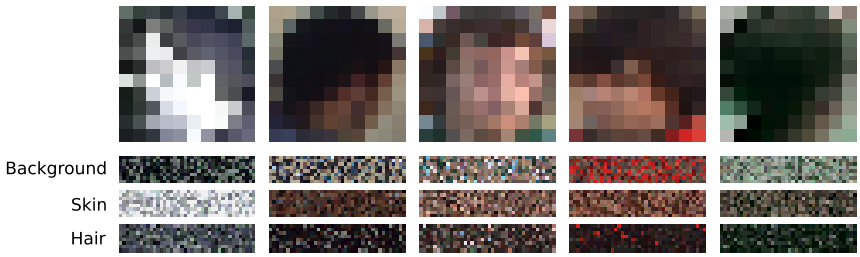


Figure 9: Five different head images and samples from the colour histograms that have been learnt up to and including the current frame

performance, twelve people were asked to estimate the pan angle of every one-hundredth head image in the same video sequences that were used to test the classifier using a graphical aid similar to figure 1. To ensure a fair comparison, the head images from the same actors were grouped to allow the test subjects to infer the hair and skin colours. The combined results show that people are capable of estimating the head pose in the test sequences with a mean absolute angular error of 26.6 degrees. Whilst the ability of people to identify head pose is fairly uniform, performance is dependant on the amount of useful information in the images, so the mean human error provides a useful measure for comparing the difficulty of different data sets.

5 Conclusion

It has been shown that segmented head images provide enough information to allow pose estimation in low resolution video, removing the need for any assumptions about hair or skin colours. The approach presented provides a high degree of invariance to lighting conditions and has the additional benefit of recovering individual hair and skin colour histograms for each of the people in the video (figure 9). These histograms could potentially be used to localise the head region in subsequent frames or to help identify the same person in situations where there are multiple cameras. The implementation achieved real-time performance on a 2.4GHz CPU with each estimation taking only ten milliseconds.

References

[1] Sileye O. Ba and Jean-Marc Odobez. Evaluation of multiple cue head pose estimation algorithms in natural environements. In *ICME*, pages 1330–1333. IEEE, 2005.

[2] Lisa M. Brown and Ying-Li Tian. Comparative study of coarse head pose estimation. In *MOTION '02: Proceedings of the Workshop on Motion and Video Computing*, page 125, Washington, DC, USA, 2002. IEEE Computer Society.

[3] J. Gonzlez, F. Xavier Roca, and J. Jos Villanueva. Hermes: A research project on human sequence evaluation. In *Computational Vision and Medical Image Processing (VipIMAGE'2007)*, 2007.

[4] Y. Huang, S. Lin, H.Q. Lu, and H.Y. Shum. Face alignment using intrinsic information. In *ICIP04*, pages V: 3307–3310, 2004.

[5] Christopher Jaynes, Amit Kale, Nathaniel Sanders, and Etienne Grossmann. The terrascope dataset: A scripted multi-camera indoor video surveillance dataset with ground-truth. In *Proceedings of the IEEE Workshop on VS PETS*, October 2005.

[6] A. Nikolaidis and I. Pitas. Facial feature extraction and pose determination. *PR*, 33(11):1783–1791, November 2000.

[7] Sourabh Niyogi and William T. Freeman. Example-based head tracking. In *FG*, pages 374–378. IEEE Computer Society, 1996.

[8] Mustafa Özuysal, Pascal Fua, and Vincent Lepetit. Fast keypoint recognition in ten lines of code. In *CVPR*. IEEE Computer Society, 2007.

[9] Ravikanth Pappu and Paul A. Beardsley. A qualitative approach to classifying gaze direction. In *FG*, pages 160–165. IEEE Computer Society, 1998.

[10] Neil Robertson and Ian Reid. Estimating gaze direction from low-resolution faces in video. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *ECCV (2)*, volume 3952 of *Lecture Notes in Computer Science*, pages 402–415. Springer, 2006.

[11] Ying-Li Tian, Lisa Brown, Jonathan H. Connell, Sharath Pankanti, Arun Hampapur, Andrew W. Senior, and Ruud M. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *AMFG*, pages 92–99. IEEE Computer Society, 2003.

[12] Antonio Torralba and Pawan Sinha. Detecting faces in impoverished images. AI Memo 2001-028, Massachusetts Institute of Technology, November 2001.

[13] Michael Voit, Kai Nickel, and Rainer Stiefelhagen. A bayesian approach for multi-view head pose estimation. *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 31–34, Sept. 2006.

[14] Ce Wang and Michael Brandstein. Robust head pose estimation by machine learning. In *ICIP*, 2000.

[15] Ying Wu and Kentaro Toyama. Wide-range, person- and illumination-insensitive head orientation estimation. In *FG*, pages 183–188. IEEE Computer Society, 2000.