5

Estimation of Head Pose

Michael Voit[1], Nicolas Gourier[2], Cristian Canton-Ferrer[3], Oswald Lanz[4], Rainer Stiefelhagen[1], Roberto Brunelli[4]

[1] Universität Karlsruhe (TH), Interactive Systems Labs, Fakultät für Informatik, Karlsruhe, Germany
[2] INRIA, Rhône-Alpes, Saint Ismier Cedex, France
[3] Universitat Politècnica de Catalunya, Barcelona, Spain
[4] Foundation Bruno Kessler, irst, Trento, Italy

In building proactive systems for interacting with users by analyzing and recognizing scenes and settings, an important task is to deal with people's occupations: Not only do their locations or identities become important, but their looking direction and orientation are crucial cues to determine everybody's intentions and actions. The understanding of interaction partners or targeted objects is relevant in deciding whether any unobtrusive system should become aware of possible matters or engaged in conversations.

The recognition of looking directions is a subtle task of either tracking people's eye gaze or finding an approximation that allows for rather unobtrusive observations, since capturing pupils requires more or less highly detailed recordings that can be made possible only with nearby standing cameras that rather limit any user's range of movement. Such an approximation can be found in estimating people's head orientation.

Whereas eye gaze allows to perceive even the smallest changes of a respective person's looking direction, the one estimation of head orientation shows its strength, especially upon low-resolution textured captures where in-depth analysis of facial features, as pupils, is not possible at all. When users are allowed to move without restrictions throughout an entire room, such a loss of detail happens quite often. Furthermore, any rotation of the head, such that only views of the back of the head or profile captures are available, makes it impossible to gather information about the user's eye gaze but still allows one to derive an albeit coarse estimation of the head's rotation and, with such, knowledge about the person's orientation. All these advantages, however, require techniques that are able to provide good generalization, especially considering the strong variance of a head's appearance, depending on the viewpoint from which the observation is made. Eventually, the dedicated head representation not only includes sharp facial features when frontal shots of heads are available but it also spans small and blurred captures as well as profile views and ambiguous shots of the back of the head, when persons are allowed to move away from the camera, which most likely is the case in intelligent environments. Compared to rather predefined single-user work spaces where a person is expected to sit in front of his or her display and a camera always delivers near-frontal observations, unrestricted movement needs to cope not only with a wider range of head rotations, and with such a stronger variance of head appearances, but also with the surrounding room's features such as the maximum distance to the camera, variance in lighting throughout the room, and occlusions.

Considering these two very different scenarios we encountered during the CHIL project, we therefore mostly distinguished between the following camera setups and expected a priori statements:

1. Single-camera settings:
   In this sensor setup, only one camera is used to classify mostly near-frontal head orientations in the range of $-90°$ to $+90°$. Usually, the person whose head orientation is to be estimated is standing or sitting right in front of that camera, thus providing face captures with rather high-resolution textures and restricted head movement.
2. Multicamera settings:
   Multicamera environments are to help overcome the limitations of single-camera settings: Observed people should have the freedom to move without boundaries – this also includes the observation of their corresponding head rotations; to cope with captures of the back of people's heads, several more cameras guarantee to capture at least one frontal shot and further profile views of the same person. This scenario not only requires state-of-the-art pose estimators, but also fusion techniques that merge multiple views or single estimates into one joint hypothesis. Head captures vary strongly in size and facial details appear mostly blurred (due to the rather high distance to the cameras) or vanish completely, as the head rotates away from a single camera's viewpoint.

In the remainder of this chapter, both sensor environments are further explored and our corresponding work in CHIL is introduced and summarized. Section 5.1 describes two techniques on the very popular topic of using only a single camera in front of a person. Section 5.2 copes with scenarios where more than one camera is available and concentrates on fusing several single-view hypotheses and joint estimation techniques. We present all approaches with their individual advantages and provide common evaluation results on INRIA's Pointing04 Database [6] for single-view head captures and on the data set of the CLEAR 2007 evaluation [11] for multiview scenarios.

5.1 Single-Camera Head Pose Estimation

Single-camera head pose estimation mostly copes with people sitting in front of a camera, showing profile or at least near-frontal face captures all the time. This leads to rather detailed captures of the user's head and face, in contrast to scenarios, where having no restrictions on people's trajectories leads to huge distances to observing cameras, thus providing only small-sized head captures where details in facial resolution are often lost or at least blurred. Most appearance-based classifiers have the ability to perform well under both circumstances since the classification is based on the whole-image representation only, no matter how detectable nostrils or lip corners are.

5.1.1 Classification with Neural Networks

A popular appearance-based approach to estimate head orientation in single views is the use of neural networks [12, 5, 10, 13]. UKA-ISL adopted this scheme under CHIL [14], where an overall accuracy of 12.3° and 12.8° could be achieved for pan and tilt estimation, respectively, on the Pointing04 database. Neural networks follow their biological counterpart and therefore mostly show their strength in generalization: After training the network on example images, this classifier has the ability to interpolate and generalize for new, unseen head images, thus allowing for almost continuous pose estimations. The network itself only receives a preprocessed head image – preprocessing usually involves the enhancement of the image's contrast to elaborate facial details – upon which the output is based. This output can either consist of a horizontal estimation only, or include further output neurons for hypothesizing the vertical orientation, too.

5.1.2 Refining Pose Estimates Using Successive Classifiers

The disadvantage of regular classifiers, even neural networks, for estimating pose is that they regularly do not allow for balancing between faster but less detailed results and deeper searches that typically deliver more accurate output. To overcome this drawback, a classification that consists of several steps, each refining the previous gathered result, is advisable. Since higher accuracy mostly goes hand in hand with higher run time, especially coarse estimates need to show a good balance between their resolution and speed. INRIA presented such a new approach in [5], where both the holistic appearance of the face as well as local details within it are combined to receive a refined classification of observed head poses. This new two-step approach performed with state-of-the-art accuracy as good as 10.1° mean error in pan and 12.6° in tilt estimation for unknown subjects.

After normalizing tracked face regions in size and slant, these normalized captures are used as a basis for a coarse estimation step by projecting them onto a *linear auto-associative memory* (LAAM), learned using the *Widrow-Hoff rule*. LAAMs are a specific case of neural networks where each input pattern is associated with each other. The Widrow-Hoff rule increases a memory's performance. At each presentation of an image, each cell of the memory modifies its weights from the others by correcting the difference between the response of the system and the desired response. A coarse head orientation estimation can then easily be made possible by searching for the prototype that best matches a current image. The advantages of using LAAMs are that they require very few parameters to be built (which allows for easy saving and reloading) but they also show robustness to partial occlusion. This allows for quite fast and easy-to-implement algorithms to gather information about a coarse pose in advance, whereas a refined classification might follow successively. Such a successive refinement can be achieved in multiple ways, by applying quite detailed classifiers to the collected face region. A trend in current research is to use wavelet families for refined estimations; however, Gaussian receptive fields proved to be less expensive than the often-used Gabor wavelets but do show interesting properties such as rotation invariance and scalability. Gaussian receptive fields motivate the construction of a model graph for each pose: Each node of the graph can be displaced locally according to its saliency in the image. These model graphs, called *salient grid graphs*, do not require a priori knowledges or heuristics about the human face and can therefore be adapted to pose estimation of other deformable objects. Whereas LAAMs only deliver a coarse estimate of the observed pose, a successive search among the coarse pose neighbors results in a final determination of the model graph that obtains the best match. The pose associated with it can then be selected as the final head orientation estimate.

5.2 Multicamera Head Pose Estimation

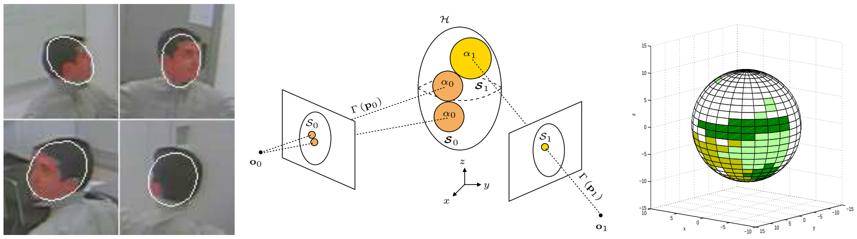
Single-camera head pose estimation has been well evaluated during years of research. Following the trend of focusing upon real-life scenarios and bringing computers into everyday living environments, that task changed to cope with observations that do not build on predefined restrictions for the user. The use of only one sensor to cover an entire room for following and tracking people's actions would never result in respective accuracies as achievable as with dedicated restrictions to only be presented with near-frontal shots of people's heads. A logical step is thus to equip a room with multiple cameras so that at least one observation always guarantees near-frontal views of the tracked person, no matter how he or she moves throughout the room. This introduces several new problems: Depending on the user's position in the room, his or her head size strongly varies over different camera observations: The nearer one stands to a camera, the bigger the head appears. The further away a person is moving, the smaller his head appears, whereas facial details vanish into blurriness or cannot be detected at all. Further issues arise around how to combine numerous views into one joint, final estimate. This *fusion* can be achieved by merging on either the signal level or a higher level. Fusing on the signal level allows for the overall dimension and processing overhead to be reduced by limiting the classification problem to one combined feature, whereas higher-level techniques often allow one to include (available) context information and help choose a smaller subset of advantageous camera views or at least leave the possibility open to extend the overall system with further sensors without the need to retrain underlying classifiers. This new task was first defined in the CHIL project and was later evaluated during the CLEAR 2006 and 2007 evaluations [14, 15, 1, 17, 16, 2, 3].

5.2.1 From Single-View Estimates to Joint Hypotheses

An approach that gathers several single-view classifiers into one successive combination and allows for an accuracy of up to 8.5° and 12.5° for horizontal and vertical orientation estimation, respectively, on the CLEAR 2007 data set can be found in UKA-ISL's publication [15], where neural networks were applied to every camera provided in CHIL smart rooms. All in all, hypothesizing over all interesting rotation directions, a single network was trained to classify camera-relative estimates: Due to the cameras' different locations, every view depicts highly different poses that need to be coped with. Training a classifier to camera-relative orientations, a successive transformation into the world coordinate system overcomes this discrepancy. That way, the classifier becomes invariant to location changes or any possible extension. For their advantage in generalization, neural networks show their strength both in single-view as well as multiview environments where face size differs strongly, as long as the training database includes sufficient examples of later observations. This single classifier can then be applied to every camera provided in the room for gathering as many single-view hypotheses of the same observed head as there are cameras. The fusion is kept independent from the classification itself: An intuitive approach is to build the average of all camera estimates into a merged output. Relying on a single neuron's output for each view, as suggested in Section 5.1.1, however, results in including a lot of noise and the overall estimate varies strongly over time. A far better way is to train the network not to output one continuous estimate, but, in fact, to describe a likelihood distribution over the defined range of possible head pose observations (i.e., $-180°$ to $+180°$ for the horizontal rotation). By further letting this distribution include the classifier's uncertainty, a successive merging of all views can be implemented by averaging the likelihood values of all single-view distributions in a Bayesian filter scheme. As described in [15], two such filters are used to track horizontal (pan) and vertical (tilt) head rotation separately. Following Bayes' rule, such a filter computes the likelihood of being in a given state, which corresponds to a certain rotation angle, depending on a previously observed likelihood distribution (the a priori knowledge) and a current measurement. Whereas the a priori knowledge implies some temporal smoothing by including the previous state distribution itself, the current measurement is obtained by building the average of all cameras' estimated likelihoods for every final pose state. The gathered a posteriori distribution over all states hence presents the joint hypothesis and allows us to classify a final pose estimate, given the current observations.

5.2.2 Fusing on the Signal Level

Fusion, of course, requires processing power to necessarily run multiple classifiers instead of only one classifier. It has the advantage of using one joint feature vector that is computed from all available views. Such a possible signal-level-based fusion technique can be found in combining spatial and color analysis to exploit redundancy, as shown by UPC in [4]. The technique presented there was also evaluated on the CLEAR 2007 data set and showed an overall accuracy of 20.48° mean error for pan estimation [3]. The system itself builds upon the idea of producing synthetic reconstructions of different head poses and searching through those templates with a new, currently achieved query vector. Since one of the face's very distinct features is the observable amount of skin, a first step in constructing the feature representation is to gather skin patches. The intuition behind this approach is that the combination of skin patches over all camera views allows for a reconstruction of skin distribution over all possible head poses. As described in [7], the probabilistic classification of pixels to contain skin color can be computed solely on their RGB information, where a distribution of skin color can be computed by means of offline hand-selected samples of skin pixels. The classification of all skin pixels in a head region and the backprojection from camera space into world coordinates then allow an ellipsoid reconstruction of skin distribution: an approximation of the head's shape and color information [see Fig. 5.1(c)]. For classifying such descriptors by matching with precomputed templates, a planar representation provides a saliency map that is easy to interpret and can be used as a likelihood evaluation function for the 3D ellipsoid's voxels and its derived head orientation. Depictions of such interpretations are shown in Fig. 5.2.



(a) (b) (c)

Fig. 5.1. Color and spatial information fusion process scheme. Ellipsoid describing the head and skin patches are detected on the original images in (a). In (b), skin pixels are backprojected onto the surface of the ellipsoid. The image in (c) shows the result of information fusion, obtaining a synthetic reconstruction of the face's appearance.

5.2.3 Integrated 3D Tracking and Pose Estimation

A common misconception of dedicated head pose estimating systems is the task of head alignment, that is, detecting an optimal head bounding box upon which the final estimation can be based by cropping this region of interest and interpreting it as a dedicated head capture. This detection is assumed to be coped with in external head tracking and alignment systems, which most often work independently of any further person tracking. Tightly linked modules might therefore provide both an increase in speed as well as better generalization, considering misaligned head regions, and overall an improvement in a tracker's accuracy due to possible head pose confidence feedback.

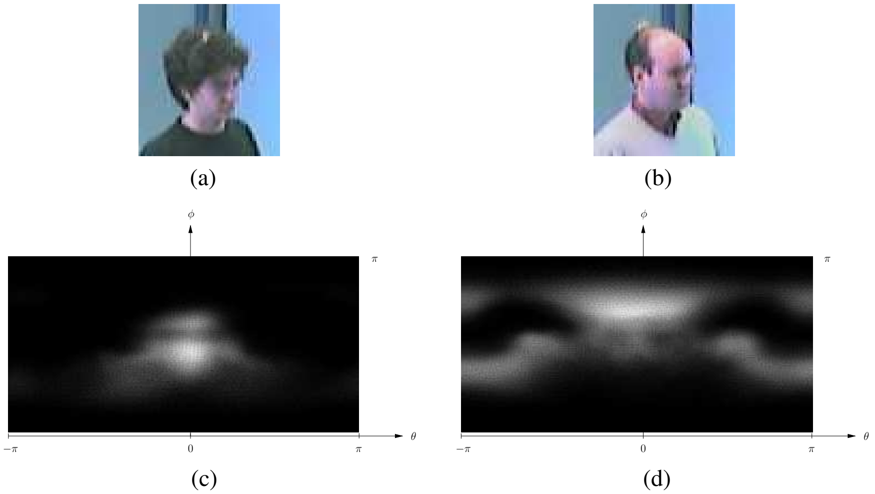

(a) (b) (c) (d)

Fig. 5.2. Planar representation of the head appearance information.

One such integration was presented by FBK-irst in [8] by means of a Bayesian estimation problem that includes both 2D body position and moving velocity as well as horizontal and vertical head orientation. In every frame step, a hypothesized body position can be updated along with its corresponding velocity component according to the time elapsed between these two frames. To account for uncertainty and ambiguity, a particle filter allows one to propagate numerous hypotheses.

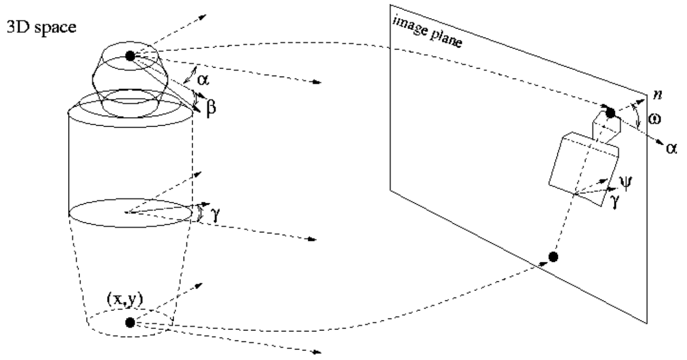


Fig. 5.3. 3D shape model parameterized by floor position (x,y), body orientation γ, and head pose (α,β) (pan and tilt angle), and an approximate, but efficient, rendering implementation that still conveys imaging artifacts such as perspective distortion and scaling. Note the offset of the head patch from the central axis; it gives a strong cue for head pose estimation. Involved in the rendering are the angular offsets ω and ψ of the body parts to the camera's optical axis n.

Low-dimensional shape and appearance models identifying image patches where the head and torso are expected to appear in each camera frame then help in computing each hypothesis' likelihood. One such adopted model is depicted in Fig. 5.3: Each hypothesis is used to construct a synthesized 3D appearance of the tracked body by assembling a set of rigid cone trunks. These trunks are positioned, scaled, and oriented according to floor location, target height, and body part orientations – a triple of 3D points, representing the centers of hips, shoulders, and top of head – that is computed from the hypothesized vector. This allows for an efficient and fast evaluation of every hypothesis. These vertices are backprojected onto every camera frame for gathering 2D segments, within which color histograms are to describe the appearance of the individual body parts. By previously collecting corresponding histograms of all body parts (which can be easily obtained upon a person's entrance into a room), potential head and upper torso patches can easily be identified within the image by comparing the corresponding histograms to their respective general counterparts. Interpolating between these templates allows for synthesizing new poses and views (Fig. 5.4). Finally, by multiplying all single-view scores, a joint multiview value can be obtained that allows one to classify for the best pose and location. Evaluated on the CLEAR 2007 data set, an overall mean error of 29.52° horizontally and 16.32° could thus be obtained [9].

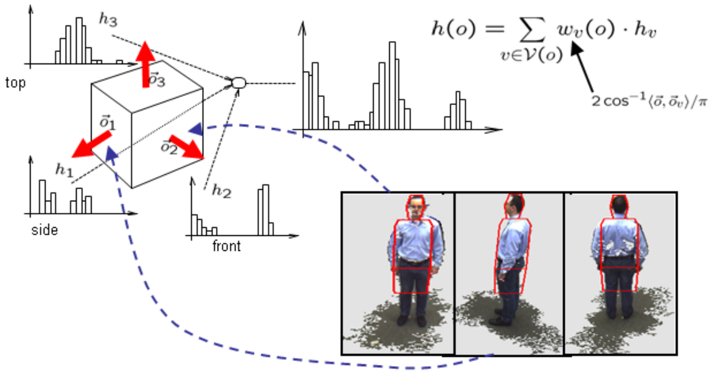

h_3
top
$\vec{o}_3$
$h(o) = \sum_{v \in \mathcal{V}(o)} w_v(o) \cdot h_v$
$2\cos^{-1}\langle \vec{o}, \vec{o}_v \rangle / \pi$
$\vec{o}_1$
$\vec{o}_2$
h_1
side
h_2
front

Fig. 5.4. Histogram synthesis. The cube faces represent preacquired top, front, and side views of the object as seen from a new viewpoint. A weight is assigned to each view according to its amount of visible area. A new object appearance is then generated as a weighted interpolation of these histograms.

5.3 Conclusion and Future Work

This chapter has presented an overview of CHIL's perceptive task to recognize people's head orientation under both monocular and multiview scenarios. Head pose estimation in single-view environments has already been the subject of numerous research projects. CHIL's main contribution in this field focused on extending that task toward using multiple cameras and allowing tracked persons to move and work without any limitations that eventually remained from a single camera's setup. During CLEAR evaluations, all developed systems were compared on publicly made data sets for both conditional tasks, which attracted a lot of interest for further, external participants. The results, achieving mean error rates as low as $10.1°$ for pan and $12.6°$ for tilt in single-view and $8.5°$ and $12.5°$, respectively, for multiview environments, demonstrated our research to be competitive and state-of-the-art. Nevertheless, remaining issues such as lighting conditions, the diversity of different hair-styles when capturing people from their back, or evaluating processing speed against a possible enhancement of accuracy by increasing the number of camera views yet remain to be coped with for further robustness and increased usability. Head pose not only allows for an indication about a person's orientation, but rather makes it possible to approximate that person's eye gaze and looking direction to successively infer a target on which he or she is focusing. Our ongoing and future research in this field will thus analyze the (visual) focus of people's attention, based on head orientation, in order to continue with the unobtrusive setup of sensors and perception (see also Chapter 9). The joint combination of further modalities to estimate the visual focus of attention will also be the subject of future research and evaluation.

References

1. S. O. Ba and J.-M. Obodez. Probabilistic head pose tracking evaluation in single and multiple camera setups. In *Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, Baltimore, MD, May 2007. Springer.
2. C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Head pose detection based on fusion of multiple viewpoints. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*. Springer, Apr. 2006.
3. C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Head orientation estimation using particle filtering in multiview scenarios. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 317–327, Baltimore, MD, May 8-11 2007.
4. C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardàs, and J. Hernando. Audiovisual head orientation estimation with particle filters in multisensor scenarios. *EURASIP Journal on Advances in Signal Processing*, 2007.
5. N. Gourier. Machine observation of the direction of human visual focus of attention, Oct. 2006. PhD thesis.
6. N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004.
7. M. Jones and J. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46:81–96, 2002.
8. O. Lanz and R. Brunelli. Dynamic head location and pose from video. In *IEEE Conference on Multisensor Fusion and Integration*, 2006.

9. O. Lanz and R. Brunelli. Joint Bayesian tracking of head location and pose from low-resolution video. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 287–296, Baltimore, MD, May 8-11 2007.
10. R. Rae and H. J. Ritter. Recognition of human head orientation based on artificial neural networks. In *IEEE Transactions on Neural Networks*, volume 9, pages 257–265, Mar. 1998.
11. R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 3–34, Baltimore, MD, May 8-11 2007.
12. R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition - ICPR 2000*, Barcelona, Sept. 2000.
13. Y.-L. Tian, L. Brown, J. Connell, S. Pankanti, A. Hampapur, A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *IEEE International Workshop on Analysis and Modeling for Face and Gestures*, Oct. 2003.
14. M. Voit, K. Nickel, and R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, Southampton, UK, Apr. 6-7 2006. Springer.
15. M. Voit, K. Nickel, and R. Stiefelhagen. Head pose estimation in single- and multi-view environments - results on the CLEAR'07 benchmarks. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 307–316, Baltimore, MD, May 8-11 2007. Springer.
16. S. Yan, Z. Zhang, Y. Fu, Y. Hu, J. Tu, and T. Huang. Learning a person-independent representation for precise 3D pose estimation. In *Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, Baltimore, MD, May 2007. Springer.
17. Z. Zhang, Y. Hu, M. Liu, and T. Huang. Head pose estimation in seminar rooms using multi-view face detectors. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*. Springer, Apr. 2006.