Head Pose Estimation in Computer Vision: A Survey

Erik Murphy-Chutorian, *Student Member, IEEE* and Mohan Manubhai Trivedi, *Senior Member, IEEE*

***Abstract*— The capacity to estimate the head pose of another person is a common human ability that presents a unique challenge for computer vision systems. Compared to face detection and recognition, which have been the primary foci of face-related vision research, identity-invariant head pose estimation has fewer rigorously evaluated systems or generic solutions. In this paper, we discuss the inherent difficulties in head pose estimation and present an organized survey describing the evolution of the field. Our discussion focuses on the advantages and disadvantages of each approach and spans 90 of the most innovative and characteristic papers that have been published on this topic. We compare these systems by focusing on their ability to estimate coarse and fine head pose, highlighting approaches that are well suited for unconstrained environments.**

***Index Terms*— Head Pose Estimation, Human Computer Interfaces, Gesture Analysis, Facial Land Marks, Face Analysis**

I. Introduction

FROM an early age, people display the ability to quickly and effortlessly interpret the orientation and movement of a human head, thereby allowing one to infer the intentions of others who are nearby and to comprehend an important nonverbal form of communication. The ease with which one accomplishes this task belies the difficulty of a problem that has challenged computational systems for decades. In a computer vision context, *head pose estimation* is the process of inferring the orientation of a human head from digital imagery. It requires a series of processing steps to transform a pixel-based representation of a head into a high-level concept of direction. Like other facial vision processing steps, an ideal head pose estimator must demonstrate invariance to a variety of image-changing factors. These factors include physical phenomena like camera distortion, projective geometry, multi-source non-Lambertian lighting, as well as biological appearance, facial expression, and the presence of accessories like glasses and hats.

Although it might seem like an explicit specification of a vision task, head pose estimation has a variety of interpretations. At the coarsest level, head pose estimation applies to algorithms that identify a head in one of a few discrete orientations, e.g., a frontal versus left/right profile view. At the fine (i.e., granular), a head pose estimate might be a continuous angular measurement across multiple Degrees of Freedom (DOF). A system that estimates only a single DOF, perhaps the left to right movement is still a head pose estimator, as is the more complex approach that estimates a full 3D orientation and position of a head, while incorporating additional degrees of freedom including movement of the facial muscles and jaw.

In the context of computer vision, head pose estimation is most commonly interpreted as the ability to infer the orientation of a person's head relative to the view of a camera. More rigorously, head pose estimation is the ability to infer the orientation of a head relative to a global coordinate system, but this subtle difference requires knowledge of the intrinsic camera parameters to undo the perceptual bias from perspective distortion. The range of head motion for an average adult male encompasses a sagittal flexion and extension (i.e., forward to backward movement of the neck) from $-60.4°$ to $69.6°$, a frontal lateral bending (i.e., right to left bending of the neck) from $-40.9°$ to $36.3°$, and a horizontal axial rotation (i.e., right to left rotation of the head) from $-79.8°$ to $75.3°$ [26]. The combination of muscular rotation and relative orientation is an often overlooked ambiguity (e.g., a profile view of a head does not look exactly the same when a camera is viewing from the side as compared to when the camera is viewing from the front and the head is turned sideways). Despite this problem, it is often assumed that the human head can be modeled as a disembodied rigid object. Under this assumption, the human head is limited to 3 degrees of freedom (DOF) in pose, which can be characterized by *pitch*, *roll*, and *yaw* angles as pictured in Fig. 1.

Head pose estimation is intrinsically linked with visual *gaze estimation*, i.e., the ability to characterize the direction and focus of a person's eyes. By itself, head pose provides a coarse indication of gaze that can be estimated in situations when the eyes of a person are not visible (like low-resolution imagery, or in the presence of eye-occluding objects like sunglasses). When the eyes are visible, head pose becomes a requirement to accurately predict gaze direction. Physiological investigations have demonstrated that a person's prediction of gaze comes from a combination of both head pose and eye direction [59]. By digitally composing images of specific eye directions on different head orientations, the authors established that an observer's interpretation of gaze is skewed in the direction of the target's head.

A graphic example of this effect was demonstrated in the nineteenth-century drawing shown in Fig. 2 [134]. In this sketch, two views of a head are presented at different orientations, but the eyes are drawn in an identical configuration in both. Glancing at this image, it is quite clear that the perceived direction of gaze is highly influenced by the pose of the head. If the head is removed entirely, and only the eyes remain, the perceived direction is similar to that in which the head is in a frontal configuration.

Based on this observation and our belief that the human gaze estimation ability is appropriately processing the visual information, we hypothesize that a passive camera sensor with no prior knowledge of the lighting conditions has insufficient information to accurately estimate the orientation of an eye without knowledge of the head orientation as well. To support this statement, consider the proportions of visible sclera (i.e., the white area), surrounding the eye. The high-contrast between the sclera and the iris are discernible from a distance, and might have evolved to facilitate gaze perception [54]. An eye direction model

Manuscript received June 5, 2007; revised November 29, 2007 and February 28, 2008; accepted March 24, 2008.

E. Murphy-Chutorian and M. M. Trivedi are with the Computer Vision and Robotics Research Laboratory, Department of Electrical and Computer Engineering, University of California, San Diego. Email: {erikmc,mtrivedi}@ucsd.edu

Digital Object Indentifier 10.1109/TPAMI.2008.106

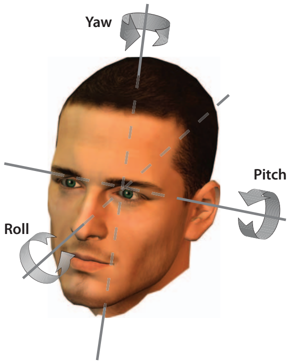

Fig. 1
THE THREE DEGREES-OF-FREEDOM OF A HUMAN HEAD CAN BE DESCRIBED BY THE EGOCENTRIC ROTATION ANGLES *pitch*, *roll*, AND *yaw*.

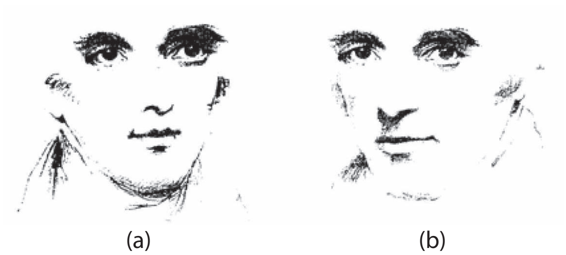

Fig. 2
WOLLASTON ILLUSION: ALTHOUGH THE EYES ARE THE SAME IN BOTH IMAGES, THE PERCEIVED GAZE DIRECTION IS DICTATED BY THE ORIENTATION OF THE HEAD [134].

that uses this scleral iris cue would need a head pose estimate to interpret the gaze direction, since any head movement introduces a gaze shift that would not affect the visible sclera. Consequently, to computationally estimate human gaze in any configuration, an eye tracker should be supplemented with a head pose estimation system.

This paper presents a survey of the head pose estimation methods and systems that have been published over the past 14 years. This work is organized by common themes and trends, paired with the discussion of the advantages and disadvantages inherent in each approach. Previous literature surveys have considered general human motion [76, 77], face detection [41, 143], face recognition [149], and affect recognition [25]. In this paper, we present a similar treatment for head pose estimation.

The remainder of the paper is structured as follows: Section II describes the motivation for head pose estimation approaches; Section III contains an organized survey of head pose estimation approaches; Section IV discusses the ground truth tools and datasets that are available for evaluation and compares the systems described in our survey based on published results and general applicability; Section V presents a summary and concluding remarks.

II. MOTIVATION

People use the orientation of their heads to convey rich, inter-personal information. For example, a person will point the direction of his head to indicate who is the intended target of a conversation. Similarly in a dialogue, head direction is a nonverbal communique that cues a listener when to switch roles and begin speaking. There is important meaning in the movement of the head as a form of gesturing in a conversation. People nod to indicate that they understand what is being said, and they use additional gestures to indicate dissent, confusion, consideration, and agreement. Exaggerated head movements are synonymous with pointing a finger, and they are a conventional way of directing someone to observe a particular location.

In addition to the information that is implied by deliberate head gestures, there is much that can be inferred by observing a person's head. For instance, quick head movements may be a sign of surprise or alarm. In people, these commonly trigger reflexive responses from an observer, which is very difficult to ignore even in the presence of contradicting auditory stimuli [58]. Other important observations can be made by establishing the visual focus of attention from a head pose estimate. If two people focus their visual attention on each other, sometimes referred to as *mutual gaze*, this is often a sign that two people are engaged in a discussion. Mutual gaze can also be used as a sign of awareness, e.g., a pedestrian will wait for a stopped automobile driver to look at him before stepping into a crosswalk. Observing a person's head direction can also provide information about the environment. If a person shifts his head towards a specific direction, there is a high likelihood that it is in the direction of an object of interest. Children as young as six months exploit this property known as *gaze following*, by looking towards the line-of-sight of a caregiver as a saliency filter for the environment [79].

Like speech recognition, which has already become entwined in many widely available technologies, head pose estimation will likely become an off-the-shelf tool to bridge the gap between humans and computers.

III. HEAD POSE ESTIMATION METHODS

It is both a challenge and our aspiration to organize all of the varied methods for head pose estimation into a single ubiquitous taxonomy. One approach we had considered was a functional taxonomy that organized each method by its operating domain. This approach would have separated methods that require stereo depth information from systems that require only monocular video. Similarly, it would have segregated approaches that require a near-field view of a person's head from those that can adapt to the low resolution of a far-field view. Another important consideration is the degree of automation that is provided by each system. Some systems estimate head pose automatically, while others assume challenging prerequisites, such as locations of facial features that must be known in advance. It is not always clear whether these requirements can be satisfied precisely with the vision algorithms that are available today.

Instead of a functional taxonomy, we have arranged each system by the fundamental approach that underlies its implementation. This organization allows us to discuss the evolution of different techniques, and it allows us to avoid ambiguities that arise when approaches are extended beyond their original functional boundaries. Our evolutionary taxonomy consists of the following eight categories that describe the conceptual approaches that have been used to estimate head pose:

TABLE I
TAXONOMY OF HEAD POSE ESTIMATION APPROACHES.

Approach	Representative Works
Appearance Template Methods	
• Image Comparison	Mean Squared Error [91], Normalized Cross-Correlation [7]
• Filtered Image Comparison	Gabor-wavelets [111]
Detector Arrays	
• Machine Learning	SVM [47], Adaboost Cascades [53]
• Neural Networks	Router Networks [103]
Nonlinear Regression Methods	
• Regression Tools	SVR [65, 86], Feature-SVR [78]
• Neural Networks	MLP [108, 115, 130], Convolutional Network [96], LLM [100], Feature-MLP [33]
Manifold Embedding Methods	
• Linear Subspaces	PCA [111], Pose-eigenspaces [114], LEA [27]
• Kernelized Subspaces	KPCA [136], KLDA [136]
• Nonlinear Subspaces	Isomap [45, 101], LE [6], LLE [102], Biased Manifold Embedding [4], SSE [142]
Flexible Models	
• Feature Descriptors	Elastic Graph Matching [55]
• Active Appearance Models	ASM [60], AAM [18, 140]
Geometric Methods	
• Facial Features	Planar & 3D Methods [30], Projective Geometry [42], Vanishing Point [132]
Tracking Methods	
• Feature Tracking	RANSAC [31, 147], Weighted Least Squares [145], Physiological Constraint [144]
• Model Tracking	Parameter Search [98, 107, 138], Least-squares [14], Dynamic Templates [139]
• Affine Transformation	Brightness and Depth Constancy [80]
• Appearance-based Particle Filters	Adaptive Diffusion [94], Dual-linear State Model [85]
Hybrid Methods	
• Geometric & Tracking	Local Feature configuration + SSD Tracking [43, 52]
• Appearance Templ. & Tracking	AVAM with Keyframes [82], Template Matching w/Particle Filtering [1]
• Nonlinear Regression & Tracking	SVR + 3D Particle Filter [85]
• Manifold Embedding & Tracking	PCA Comparison + Continuous Density HMM [49]
• Other Combinations	Nonlinear regression + Geometric + Particle Filter [109], KLDA + EGM [136]

- **Appearance Template Methods** compare a new image of a head to a set of exemplars (each labeled with a discrete pose) in order to find the most similar view.
- **Detector Array Methods** train a series of head detectors each attuned to a specific pose and assign a discrete pose to the detector with the greatest support.
- **Nonlinear Regression Methods** use nonlinear regression tools to develop a functional mapping from the image or feature data to a head pose measurement.
- **Manifold Embedding Methods** seek low-dimensional manifolds that model the continuous variation in head pose. New images can be embedded into these manifolds and then used for embedded template matching or regression.
- **Flexible Models** fit a non-rigid model to the facial structure of each individual in the image plane. Head pose is estimated from feature-level comparisons or from the instantiation of the model parameters.
- **Geometric Methods** use the location of features such as the eyes, mouth, and nose tip to determine pose from their relative configuration.
- **Tracking Methods** recover the global pose change of the head from the observed movement between video frames.
- **Hybrid Methods** combine one or more of these aforementioned methods to overcome the limitations inherent in any single approach.

Table I provides a list of representative systems for each of these categories. In this section, each category is described in detail. Comments are provided on the functional requirements of each approach and the advantages and disadvantages of each design choice.

A. Appearance Template Methods

Appearance template methods use image-based comparison metrics to match a view of a person's head to a set of exemplars with corresponding pose labels. In the simplest implementation, the queried image is given the same pose that is assigned to the most similar of these templates. An illustration is presented in Fig. 3. Some characteristic examples include the use of normalized cross-correlation at multiple image resolutions [7] and mean squared error (MSE) over a sliding window [91].

Appearance templates have some advantages over more complicated methods. The templates can be expanded to a larger set at any time, allowing systems to adapt to changing conditions. Furthermore, appearance templates do not require negative training examples or facial feature points. Creating a corpus of training data requires only cropping head images and providing head pose annotations. Appearance templates are also well suited for both high and low-resolution imagery.

There are many disadvantages with appearance templates. Without the use of some interpolation method, they are only capable of estimating discrete pose locations. They typically assume that the head region has already been detected and localized, and localization error can degrade the accuracy of the head pose estimate. They can also suffer from efficiency concerns, since as more templates are added to the exemplar set, more computationally expensive image comparisons will need to be computed. One proposed solution to these last two problems was to train a set of Support Vector Machines (SVMs) to detect and localize the face, and subsequently use the support vectors as appearance templates to estimate head pose [88, 89].

Despite those limitations, the most significant problem with appearance templates is that they operate under the faulty assumption that pairwise similarity in the image space can be equated to similarity in pose. Consider two images of the same person

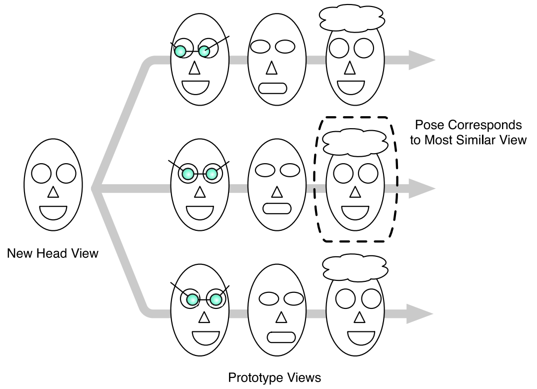

Fig. 3

APPEARANCE TEMPLATE METHODS COMPARE A NEW HEAD VIEW TO A SET OF TRAINING EXAMPLES (EACH LABELED WITH A DISCRETE POSE) AND FIND THE MOST SIMILAR VIEW.

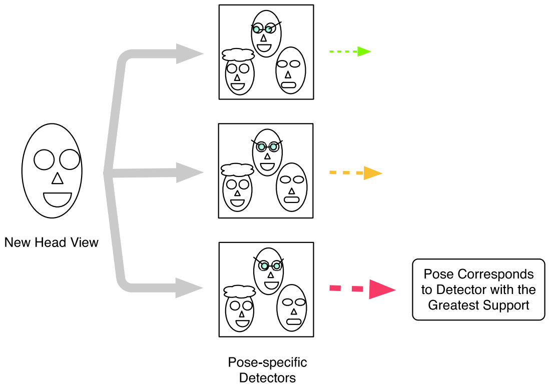

Fig. 4

DETECTOR ARRAYS COMPRISE A SERIES OF HEAD DETECTORS, EACH ATTUNED TO A SPECIFIC POSE, AND A DISCRETE POSE IS ASSIGNED TO THE DETECTOR WITH THE GREATEST SUPPORT.

at slightly different poses and two images of different people at the same pose. In this scenario, the effect of identity can cause more dissimilarity in the image than from a change in pose, and template matching would improperly associate the image with the incorrect pose. Although this effect would likely lessen for widely varying poses, there is still no guarantee that pairwise similarity corresponds to similarity in the pose domain (e.g., a right-profile image of a face may be more similar to a left-profile image than to a frontal view). Thus, even with a homogeneous set of discrete poses for every individual, errors in template comparison can lead to highly erroneous pose estimates.

To decrease the effect of the pairwise similarity problem, many approaches have experimented with various distance metrics and image transformations that reduce the head pose estimation error. For example, the images can be convolved with a Laplacian-of-Gaussian filter [32] to emphasize some of the more common facial contours while removing some of the identity-specific texture variation across different individuals. Similarly, the images can be convolved with a complex Gabor-wavelet to emphasize directed features, such as the vertical lines of the nose and horizontal orientation of the mouth [110, 111]. The magnitude of this complex convolution also provides some invariance to shift, which can conceivably reduce the appearance errors that arise from the variance in facial feature locations between people.

B. Detector Arrays

Many methods have been introduced in the last decade for frontal face detection [97, 104, 126]. Given the success of these approaches, it seems like a natural extension to estimate head pose by training multiple face detectors, each to specific a different discrete pose. Fig. 4 illustrates this approach. For arrays of binary classifiers, successfully detecting the face will specify the pose of the head, assuming that no two classifiers are in disagreement. For detectors with continuous output, pose can be estimated by the detector with the greatest support. Detector arrays are similar to appearance templates in that they operate directly on an image patch. Instead of comparing an image to a large set of individual templates, the image is evaluated by a detector trained on many images with a supervised learning algorithm.

An early example of a detector array used three SVMs for three discrete yaws [47]. A more recent system trained five FloatBoost classifiers operating in a far-field, multi-camera setting [146].

An advantage of detector array methods is that a separate head detection and localization step is not required, since each detector is also capable of making the distinction between head and non-head. Simultaneous detection and pose estimation can be performed by applying the detector to many subregions in the image. Another improvement is that unlike appearance templates, detector arrays employ training algorithms that learn to ignore the appearance variation that does not correspond to pose change. Detector arrays are also well suited for high and low-resolution images.

Disadvantages to detector array methods also exist. It is burdensome to train many detectors for each discrete pose. For a detector array to function as a head detector and pose estimator, it must also be trained on many negative non-face examples, which require substantially more training data. In addition, there may be systematic problems that arise as the number of detectors is increased. If two detectors are attuned to very similar poses, the images that are positive training examples for one must be negative training examples for another. It is not clear whether prominent detection approaches can learn a successful model when the positive and negative examples are very similar in appearance. Indeed, in practice these systems have been limited to one degree of freedom and fewer than 12 detectors. Furthermore, since the majority of the detectors have binary outputs, there is no way to derive a reliable continuous estimate from the result, allowing only coarse head pose estimates and creating ambiguities when multiple detectors simultaneously classify a positive image. Finally, the computational requirements increase linearly with the number of detectors, making it difficult to implement a real-time system with a large array. As a solution to this last problem, it has been suggested that a *router* classifier can be used to pick a single subsequent detector to use for pose estimation [103]. In this case the router effectively determines pose (i.e., assuming this is a face, what is its pose?), and the subsequent detector confirms the choice (i.e., is this a face at the pose specified by the router?). Although this technique sounds promising in theory, it should be noted that it has not been demonstrated for pitch or yaw change, but rather only for rotation in the camera-plane, first using neural network-based face detectors [103], and later with cascaded AdaBoost detectors [53].

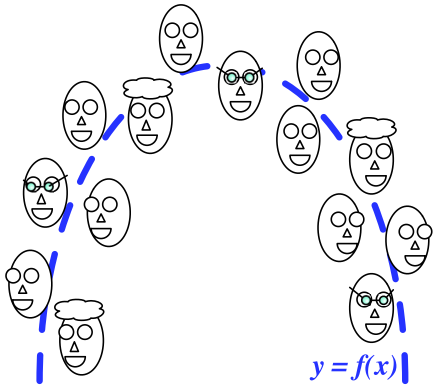

Fig. 5

Nonlinear regression provides a functional mapping from the image or feature data to a head pose measurement.

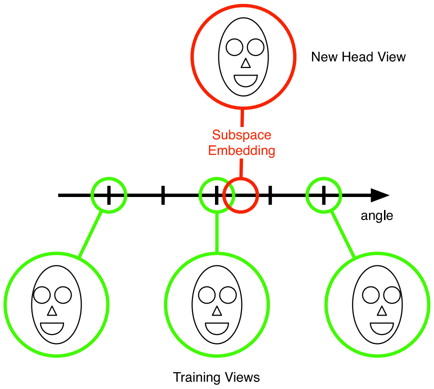

Fig. 6

Manifold Embedding Methods try to directly project a processed image onto the head pose manifold using linear and non-linear subspace techniques.

C. Nonlinear Regression Methods

Nonlinear regression methods estimate pose by learning a nonlinear functional mapping from the image space to one or more pose directions. An illustration is provided in Fig. 5. The allure of these approaches is that with a set of labeled training data, a model can be built that will provide a discrete or continuous pose estimate for any new data sample. The caveat with these approaches is that it is not clear how well a specific regression tool will be able to learn the proper mapping.

The high-dimensionality of an image presents a challenge for some regression tools. Success has been demonstrated using Support Vector Regressors (SVRs) if the dimensionality of the data can be reduced, as for example with Principal Component Analysis (PCA) [64, 65], or with localized gradient orientation histograms [86] – the latter giving better accuracy for head pose estimation. Alternatively, if the location of facial features are known in advance, regression tools can be used on relatively low-dimensional feature data extracted at these points [70, 78].

Of the nonlinear regression tools used for head pose estimation, neural networks have been the most widely used in the literature. An example is the multi-layer perceptron (MLP), consisting of many feed-forward cells defined in multiple layers (e.g. the output of the cells in one layer comprise the input for the subsequent layer) [8, 23]. The perceptron can be trained by backpropagation, which is a supervised learning procedure that propagates the error backwards through each of the hidden layers in the network to update the weights and biases. Head pose can be estimated from cropped images of a head using an MLP in different configurations. For instance, the output nodes can correspond to discretized poses, in which case a Gaussian kernel can be used to smooth the training poses to account for the similarity of nearby poses [10, 106, 148]. Although this approach works, it is similar to detector arrays and appearance templates in that it provides only a coarse estimate of pose at discrete locations.

An MLP can also be trained for fine head pose estimation over a continuous pose range. In this configuration, the network has one output for each DOF, and the activation of the output is proportional to its corresponding orientation [108, 115, 117, 129, 130]. Alternatively, a set of MLP networks with a single output node can be trained individually for each DOF. This approach has been used for heads viewed from multiple far-field cameras in indoor environments, using background subtraction or a color filter to detect the facial region and Bayesian filtering to fuse and smooth the estimates of each individual camera [120, 128–130].

A locally-linear map (LLM) is another popular neural network consisting of many linear maps [100]. To build the network, the input data is compared to a centroid sample for each map and used to learn a weight matrix. Head pose estimation requires a nearest-neighbor search for the closest centroid, followed by linear regression with the corresponding map. This approach can be extended with difference vectors and dimensionality reduction [11] as well as decomposition with Gabor-wavelets [56].

As was mentioned earlier with SVR, it is possible to train a neural network using data from facial feature locations. This approach has been evaluated with associative neural networks [33, 34].

The advantages of neural network approaches are numerous. These systems are very fast, only require cropped labeled faces for training, work well in near-field and far-field imagery, and give some of the most accurate head pose estimates in practice (see Section IV).

The main disadvantage to these methods is that they are to prone to error from poor head localization. As a suggested solution, a convolutional network [62] that extends the MLP by explicitly modeling some shift, scale, and distortion invariance can be used to reduce this source of error [95, 96].

D. Manifold Embedding Methods

Although an image of a head can be considered a data sample in high-dimensional space, there are inherently many fewer dimensions in which pose can vary. With a rigid model of the head, this can be as few as three dimensions for orientation and three for position. Therefore, it is possible to consider that each high-dimensional image sample lies on a low-dimensional continuous manifold constrained by the allowable pose variations. For head pose estimation, the manifold must be modeled, and an embedding technique is required to project a new sample into the manifold. This low-dimensional embedding can then be used for head pose estimation with techniques such as regression in the embedded space or embedded template matching. Any dimensionality reduction algorithm can be considered an attempt at manifold embedding, but the challenge lies in creating an algorithm that successfully recovers head pose while ignoring other sources of image variation.

Two of the most popular dimensionality reduction techniques, principal component analysis (PCA) and its nonlinear kernelized version KPCA, discover the primary modes of variation from a set of data samples [23]. Head pose can be estimated with PCA, e.g., by projecting an image into a PCA subspace and comparing the results to a set of embedded templates [75]. It has been shown that similarity in this low-dimensional space is more likely to correlate with pose similarity than appearance template matching with Gabor-wavelet preprocessing [110, 111]. Nevertheless, PCA and KPCA are inferior techniques for head pose estimation [136]. Besides the linear limitations of standard PCA that cannot adequately represent the nonlinear image variations caused by pose change, these approaches are unsupervised techniques that do not incorporate the pose labels that are usually available during training. As a result, there is no guarantee that the primary components will relate to pose variation rather than to appearance variation. Probably, they will be correspond to both.

To mitigate these problems, the appearance information can be decoupled from the pose by splitting the training data into groups that each share the same discrete head pose. Then, PCA and KPCA can be applied to generate a separate projection matrix for each group. These pose-specific eigenspaces, or *pose-eigenspaces*, each represent the primary modes of appearance variation and provide a decomposition that is independent of the pose variation. Head pose can be estimated by normalizing the image and projecting it into each of the pose-eigenspaces, thus finding the pose with the highest projection energy [114]. Alternatively, the embedded samples can be used as the input to a set of classifiers, such as multi-class SVMs [63]. As testament to the limitations of KPCA, however, it has been shown that by skipping the KPCA projection altogether and using local Gabor binary patterns, one can greatly improve pose estimation with a set of multi-class SVMs [69]. Pose-eigenspaces have an unfortunate side-effect. The ability to estimate fine head pose is lost since, like detector arrays, the estimate is derived from a discrete set of measurements. If only coarse head pose estimation is desired, it is be better to use multi-class linear discriminant analysis (LDA) or the kernelized version, KLDA [23] since these techniques can be used to find the modes of variation in the data that best account for the differences between discrete pose classes [15, 136].

Other manifold embedding approaches have shown more promise for head pose estimation. These include Isometric feature mapping (Isomap) [101, 119], Locally Linear Embedding (LLE) [102], and Laplacian Eigenmaps (LE) [6]. To estimate head pose with any of these techniques, there must be a procedure to embed a new data sample into an existing manifold. Raytchev et al. [101] described such a procedure for an Isomap manifold, but for out-of-sample embedding in an LLE and LE manifold there has been no explicit solution. For these approaches, a new sample must be embedded with an approximate technique, such as a Generalized Regression Neural Network [4]. Alternatively, LLE and LE can be replaced by their linear approximations, locally embedded analysis (LEA) [27] and locality preserving projections (LPP) [39].

There are still some remaining weaknesses in the manifold embedding approaches mentioned thus far. With the exception of LDA and KLDA, each of these techniques operates in an unsupervised fashion, ignoring the pose labels that might be available during training. As a result, they have the tendency to build manifolds for identity as well as pose [4]. As one solution to this problem, identity can be separated from pose by creating a separate manifold for each subject that can be aligned together. For example, a high-dimensional ellipse can be fit to the data in a set of Isomap manifolds and then used to normalize the manifolds [45]. To map from the feature space to the embedded space, nonlinear interpolation can be performed with radial basis functions. Nevertheless, even this approach has its weaknesses, because appearance variation can be due to factors other than identity and pose, such as lighting. For a more general solution, instead of making separate manifolds for each variation, a single manifold can be created that uses a distance metric that is biased towards samples with smaller pose differences [4]. This change was shown to improve the head pose estimation performance of Isomap, LLE, and LE.

Another difficulty to consider is the heterogeneity of the training data that is common in many real-world training scenarios. To model identity, multiple people are needed to train a manifold, but it is often impossible to obtain a regular sampling of poses from each individual. Instead, the training images comprise a disjoint set of poses for each person sampled from some continuous measurement device. A proposed remedy to this problem is to create individualize submanifolds for each subject, and use them to render virtual reconstructions of the discrete poses that are missing between subjects [142]. This work introduced Synchronized Submanifold Embedding (SSE), a linear embedding that creates a projection matrix that minimizes the distances between each sample and its nearest reconstructed neighbors (based on the pose label), while maximizing the distances between samples from the same subject.

All of the manifold embedding techniques described in this section are linear or nonlinear approaches. The linear techniques have the advantage that embedding can be performed by matrix multiplication, but they lack the representational ability of the nonlinear techniques. As a middle ground between these approaches, the global head pose manifold can be approximated by a set of localized linear manifolds. This has been demonstrated for head pose estimation with PCA, LDA, and LPP [66].

E. Flexible Models

The previous methods have considered head pose estimation as a signal detection problem, mapping a rectangular region of image pixels to a specific pose orientation. Flexible models take a different approach. With these techniques, a non-rigid model is fit to the image such that it conforms to the facial structure of each individual. In addition to pose labels, these methods require training data with annotated facial features, but it enables them to make comparisons at the feature level rather than at the global appearance level. A conceptual illustration is presented in Fig. 7.

Recall the appearance template methods from Section III-A. To estimate pose, the view of a new head is overlaid on each template and a pixel-based metric is used to compare the images. Even with perfect registration, however, the images of two different people will not line up exactly, since the location of facial features vary between people. Now, consider a template based on a deformable graph of local feature points (eye corners, nose, mouth corners, etc). To train this system, facial feature locations are manually labeled in each training image, and local feature descriptors such as Gabor-jets, can be extracted at each location. These features can be extracted from views of multiple people, and extra invariance can achieved by storing a *bunch* of

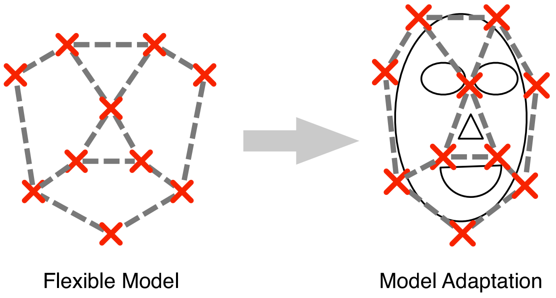

Fig. 7
FLEXIBLE MODELS ARE FIT TO THE FACIAL STRUCTURE OF THE INDIVIDUAL IN THE IMAGE PLANE. HEAD POSE IS ESTIMATED FROM FEATURE-LEVEL COMPARISONS OR FROM THE INSTANTIATION OF THE MODEL PARAMETERS.

descriptors at each node. This representation has been called an Elastic Bunch Graph [57], and possesses the ability to represent non-rigid, or deformable, objects. To compare a bunch graph to a new face image, the graph is placed over the image, and exhaustively or iteratively deformed to find the minimum distance between the features at every graph node location. This process is called Elastic Graph Matching (EGM). For head pose estimation, a different bunch graph is created for every discrete pose, and each of these are compared to a new view of the head. The bunch graph with the maximum similarity assigns a discrete pose to the head [55, 136]. Because EGM uses features located at specific facial points, there is significantly less inter-subject variability than with unaligned points. This makes it much more likely that similarity between the models will equate to similarity in pose. A disadvantage to this method is that the pose estimate is discrete, requiring many bunch graphs to gain fine head pose estimates. Unfortunately comparing many bunch graphs, each with many deformations, is computationally expensive in comparison to most other head pose estimation techniques.

Another flexible model that has evolved for head pose estimation is the Active Appearance Model (AAM) [19], which learns the primary modes of variation in facial shape and texture from a 2D perspective. Consider a set of M specific facial points, (perhaps the corners of the eyes, ear tips, nostrils, chin, and mouth). Each point has a 2D coordinate in an image, and these points can be ordered by facial feature and concatenated into a vector of length $2M$. If these feature vectors are computed for many faces, spanning the different individuals and poses in which all of the features can be found, they can be used to find the variation in facial shape. Using a dimensionality reduction technique such as PCA on this data results in an Active Shape Model (ASM) [17], capable of representing the primary modes of shape variation. Simply by looking at the largest principal components, one can find the directions in the data that correspond to variations in pitch and yaw [60, 61]. If the location of the facial features were known in a new image, pose could be estimated by projecting the feature locations into the shape subspace and evaluating the components responsible for pose. This can be accomplished by augmenting the ASM with texture information and performing an iterative search to fit the shape to a new image of a face. Early work extracted local grayscale profiles at each feature point and employed a *greedy* search to match the feature points [60, 61]. Later, the joint shape and texture AAM were introduced [19].

To build an AAM, first an ASM must be generated from a set of training data. Next, the face images must be warped such that the feature points match those of the mean shape. The warped images should be normalized, and then used to build a shape-free texture model, (originally a texture-based PCA subspace). Finally, the correlations between shape and texture are learned and used to generate a combined appearance (shape and texture) model [24]. Given a rough initialization of face shape, the AAM can be fit to a new facial image by iteratively comparing the rendered appearance model to the observed image and adjusting the model parameters to minimize a distance measure between these two images. Once the model has converged to the feature locations, an estimate of head pose can be obtained by mapping the appearance parameters to a pose estimate, a simple example being yaw estimation with linear regression [18].

AAMs have come a long way since their original inception. Fitting methods based on the inverse compositional image alignment algorithm overcome the linear assumption of how appearance error relates to the gradient descent search and allows for more accurate, real-time convergence [72]. A tracked AAM over a video sequence can also be used to estimate the 3D shape modes, which can subsequently be reintroduced to constrain the 2D AAM fitting process [140]. Once the 3D constraint is learned, the AAM can be used to directly estimate the 3D orientation of the head. Alternatively, since the AAM shape points have a one-to-one correspondence, Structure From Motion (SFM) algorithms can be used to estimate the 3D shape of the face, as well as the relative pose difference between two video frames [35]. Further work with AAMs have introduced modifications that expand their utility to driver head pose estimation [3] and multiple cameras [44].

AAMs have good invariance to head localization error, since they adapt to the image and find the exact location of the facial features. This allows for precise and accurate head pose estimation. The main limitation of AAMs is that all of the facial features are required to be located in each image frame. In practice, these approaches are limited to head pose orientations from which the outer corners of both eyes are visible. It is also not evident that AAM fitting algorithms could successfully operate for far-field head pose estimation with low-resolution facial images.

F. Geometric Methods

There is a great divide between most of the computer vision pose estimation approaches and the results of psychophysical experiments. While the former have focused predominantly on appearance-based solutions, the latter consider the human perception of head pose to rely on cues such as the deviation of the nose angle and the deviation of the head from bilateral symmetry [133]. These effects and other factors, such as the location of the face in relation to the contour of the head, strongly influence the human perception of head pose, suggesting that these are extremely salient cues regarding the orientation of the head. Geometric approaches to head pose estimation use head shape and the precise configuration of local features to estimate pose, as depicted in Fig. 8. These methods are particularly interesting because they can directly exploit properties which are known to influence human head pose estimation.

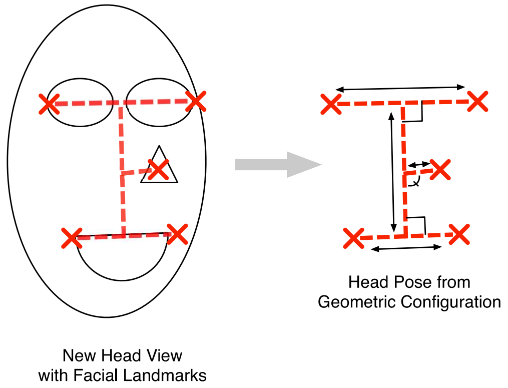

Fig. 8
Geometric Methods attempt to find local features such as the eyes, mouth, and nose tip and determine pose from the relative configuration of these features.

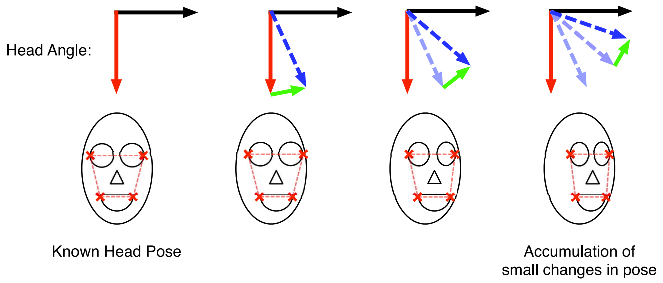

Fig. 9
Tracking Methods find the relative movement between video frames to estimate the global movement of a head.

Early approaches focused on estimating the head from a set of facial feature locations. It is assumed that these features are already known, and that pose can be estimated directly from the configuration of these points.

The configuration of facial features can be exploited in many ways to estimate pose. Using five facial points (the outside corners of each eye, the outside corners of the mouth, and the tip of the nose) the facial symmetry axis is found by connecting a line between the midpoint of the eyes and midpoint of the mouth [30]. Assuming a fixed ratio between these facial points and a fixed length of the nose, the facial direction can be determined under weak-perspective geometry from the 3D angle of the nose. Alternatively, the same five points can be used to determine the head pose from the normal to the plane, which can be found from planar skew-symmetry and a coarse estimate of the nose position [30]. Another estimate of pose can be obtained using a different set of 5 points (the inner and outer corners of each eye, and the tip of the nose) [42]. Under the assumption that all four eye points are assumed to be coplanar, yaw can be determined from the observable difference in size between the left and right eye due to projective distortion from the known camera parameters. Roll can be found simply from the angle of this line from the horizon. Pitch is determined by comparing the distance between the nose tip and the eye-line to an anthropometric model. Unlike the previous two approaches, however, this technique does not present a solution to improve pose estimation for near-frontal views. These configurations are called *degenerative angles*, since they require very high precision to accurately estimate head pose with this model. Another method was recently proposed using the inner and outer corners of each eye and the corners of the mouth, which are automatically detected in the image [132]. The observation is that three lines between the outer eye corners, the inner eye corners, and the mouth are parallel. Any observed deviation from parallel in the image plane is a result of perspective distortion. The vanishing point (i.e., where these lines would intersect in the image plane) can be calculated using least squares to minimize the over-determined solution for three-lines. This point can be used to estimate the 3D orientation of the parallel lines if the ratio of their lengths is known, and it can be used to estimate the absolute 3D position of each feature point, if the actual line length is known. As this information varies across identity, the EM algorithm with a Gaussian mixture model can adapt the facial parameters for each person to minimize the backprojection error. The downside to this approach is that pose can only be estimated if the pose is near enough to a frontal view to see all of the facial lines.

These geometric methods are fast and simple. With only a few facial features, a decent estimate of head pose can be obtained. The obvious difficulty lies in detecting the features with high precision and accuracy, but the more subtle challenges stem from handling outlying or missing feature detections. Far-field imagery, in this context, is problematic since the resolution can make it difficult or impossible to precisely determine the feature locations. Also, situations often arise that can permanently obscure facial landmarks, such as when a person wears glasses and obscures his eye corners. Considering that geometric approaches depend on the accurate detection of facial points, they are typically more sensitive to occlusion than appearance-based methods that use information from the entire facial region.

It is worth noting that even very simple geometric cues can be used to estimate head pose. Fitting an ellipse to the gradient contour and color of a face provides a coarse estimate of pose for one degree of freedom [20]. For near-frontal faces, the yaw of a face can be reliably estimated by creating a triangle between the eyes and the mouth and finding its deviation from a pure isosceles triangle [90]. With multiple-cameras surrounding the head, yaw can be estimated as the orientation with the most skin color [12] or with a skin-color template [13]. Likewise, the location of the face can be estimated by effectively registering the location of the face relative to the segmented facial region [141].

G. *Tracking Methods*

Tracking methods operate by following the relative movement of the head between consecutive frames of a video sequence as depicted in Fig. 9. Temporal continuity and smooth motion constraints are utilized to provide a visually appealing estimate of pose over time. These systems typically demonstrate a high level of accuracy (see Section IV), but initialization from a known head position is requisite. Typically, the subject must maintain a frontal pose before the system has begun, and must be reinitialized whenever the track is lost. As a result, approaches often rely on manual initialization or a camera view such that the subject's neutral head pose is forward-looking and easily reinitialized with a frontal face detector.

Tracking methods can operate in a bottom-up manner, following low-level facial landmarks from frame to frame. Early work considered six feature points (tracked using correlation windows) and determined the head movement from weak-perspective geometry [31]. A more sophisticated approach is to assume the human face is a planar surface in an orthographic space. In this case, two degrees-of-freedom can be recovered by using weighted least-squares to determine the best affine transformation between any two frames. The problem is reduced to a rotational ambiguity, which is capable of providing the head direction [145]. Earlier approaches used traditional least squares to fit automatically selected facial points under affine geometry [73] and weak-perspective geometry [121]. A global SSD tracker coarsely followed the entire face, as local features were tracked from within this region. More recently, these methods have evolved into more complex techniques that match feature points with robust SIFT [68] descriptors and use prior knowledge of 3D face shape [93, 144] or stereo and RANSAC-based matching [147] to recover the pose change under full perspective projection.

Tracking can alternatively employ a model-based approach, by finding the transformation of a model that best accounts for the observed movement of the head. For head pose estimation, it is common to use a rigid 3D model of the head. To estimate head pose, one simply needs to find the rotation and translation of the model that best fits each new image-based observation. This can be accomplished by texture-mapping the image of the head onto a 3D model. In the simplest implementation, this can be done manually, and then head pose can be estimated by searching through a discrete set of transformations to find the one that minimizes the difference in appearance between the new frame and the model [98]. This can be improved to continuous pose measurement using a gradient descent search [107] and further refined with optical flow to guide the optimization [71]. Also, since a global appearance model can suffer when dynamic lighting introduces effects such as partial shadowing, a similarity metric can be used that averages over a set of localized regions [138].

Often, reasonable accuracy can be obtained with affine transformations (e.g., with a stereo camera rig, one can find the relative pose change as the translation and rotation that minimize the error in a least-squares sense for both grayscale intensity and depth [37, 80]). A similar approach has been presented with a time-of-flight sensor using constancy of 2D motion fields and depth from optical flow [150].

The primary advantage of tracking approaches is their ability to track the head with high accuracy by discovering the small pose shifts between video frames. In this tracking configuration, these methods consistently outperform other head pose estimation approaches (see Section IV). An additional advantage with model-based tracking is the ability to dynamically construct an individualized archetype of a person's head. This allows these approaches to avoid the detrimental effects of appearance variation.

The difficulty with tracking methods is the requisite of an accurate initialization of position and pose to generate a new model or adapt an existing model. Without a separate localization and head pose estimation step, these approaches can only be used to discover the relative transformation between frames. In this mode of operation, these methods are not estimating head pose in the absolute sense, but rather tracking the movement of the head. Nevertheless, for some applications, only relative motion is required. Some examples include tracking the head with a manually initialized cylindrical model and recursive least squares optimization [14], or by tracking with a deformable anthropomorphic 3D model [21]. Tracking approaches can be initialized automatically, using dynamic templates to recreate the model whenever the head pose estimate is near the original view [139].

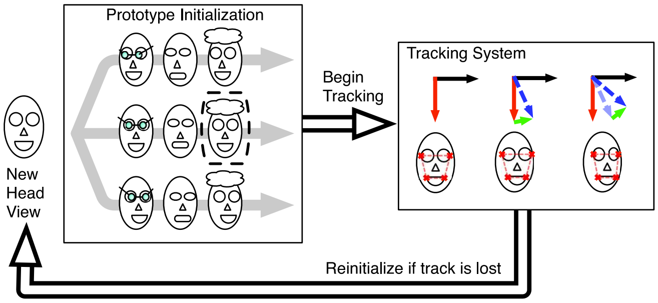

Fig. 10

Hybrid Methods combine one or more approaches to estimate pose. This image is an example of an appearance template method combined with a point tracking system.

These model-tracking approaches can be improved with appearance-based particle filtering [38, 51] to incorporate prior information about the dynamics of the head. In a classical particle filter, an observation of the object's state is observed at every time-step and assumed to be noisy; the optimal track can be found by maximizing the posterior probability of the movement given the observation, using a simulated set of samples. For appearance-based particle filtering, instead of observing a noisy sample of the head's absolute position and orientation, an image of the head is obtained at every time-step. The observation noise is negligible, and the difficulty instead lies in inferring the object's state from the image pixels. This appearance-based filtering problem can be solved with a similar construction. A set of pose samples are generated with a prior dynamic model and used to render views of the model with different transformations. Each virtual image can be directly compared to the observed image, and these comparisons can be used to update the particle filtering weights. This technique allows for accurate real-time head pose tracking in a variety of environments, including near-field video [22], low-resolution video with adaptive PCA subspaces [124], near-field stereo with affine approximations [94], and daytime and nighttime driving video with a dual-linear state model [85].

H. *Hybrid Methods*

Hybrid approaches combine one or more of the aforementioned methods to estimate pose, as illustrated by example in Fig. 10. These systems can be designed to overcome the limitations of any one specific head pose category. A common embodiment is to supplement a static head pose estimation approach with a tracking system. The static system is responsible for initialization and the tracking system is responsible for maintaining pose estimates over time. If the tracker begins to drift, the static system can reinitialize the track. This method yields the high accuracy of pure tracking approaches without initialization and drift limitations. Many successful combinations have been presented by mixing an automatic geometric method with point tracking [40, 43, 46, 52, 87], PCA embedded template matching with optical flow [151], PCA embedded template matching with a continuous density hidden Markov model [49], PCA embedded template keyframe matching with stereo tracking by grayscale and depth constancy [81], and color and texture appearance templates with image-based particle filtering [1].

Some notable extensions to these works have also been presented. The work by Morency et al. [81] was extended with pose-based eigenspaces to synthetically generate new templates from a single initialization [82]. Ba and Obodez [1] later extended and refined their technique for multiple cameras and far-field imagery [2].

Hybrid systems can also use two or more independent techniques and fuse the estimates from each system into a single result. In this case, the system gains information from multiple cues that together increase the estimation accuracy. Specific examples include appearance template matching with geometric cues (also with a particle filter) [109] and manifold embedding estimation refined by Elastic Graph Matching [136, 137].

IV. Head Pose Estimation Comparisons

A. *Ground Truth Datasets*

To evaluate and compare head pose estimation systems, an accurate method is requisite to measure the ground truth for a set of evaluative data. Typically, ground truth is essential to train any head pose estimation approach. The following list describes the most common methods for capturing this ground truth data, in an approximate order from least accurate (coarse) to most accurate (fine).

Directional Suggestion – A set of markers are placed at discrete locations around a room and each human subject is asked to direct his head towards each of the locations while a camera captures discrete images for every location [33]. This method is a poor source of ground truth. Firstly, it assumes that each subject's head is in the exact same physical location in 3D space, such that the same head directions correspond to the same pose orientations. Secondly and most importantly, it assumes that a person has the ability to accurately direct their head towards an object. Unfortunately, this is a subjective task that people tend to perform rather poorly. For instance, subjective errors can be seen in the widely-used Pointing '04 dataset [125].

Directional Suggestion with Laser Pointer – This approach is identical to directional suggestion, but a laser pointer is affixed to the subject's head [100]. This allows the subject to pinpoint the discrete locations in the room with much higher accuracy from visual feedback, but it still assumes that the subjects' heads are located at the same point in space such that direction is equivalent to head pose. In practice, this is difficult to ensure, since people have a tendency to shift their head position during data capture.

Manual Annotation – Images of human faces are viewed by a person who assigns a pose label based on his own perception of the pose. For a coarse set of poses in 1 DOF this may be sufficient, but it is inappropriate for fine head pose estimation.

Camera Arrays – In this approach, multiple cameras at known positions simultaneously capture images of a person's face from different angles. If care is taken to ensure that each subject's head is in the same location during capture, this method provides a highly accurate set of ground truth. The disadvantage is that this is only applicable for near-field images and cannot be applied to fine poses or real-world video.

Magnetic Sensors – Magnetic Sensors, such as the Polhemus *FastTrak* or Ascension *Flock of Birds*, work by emitting and measuring a magnetic field. The sensor can be affixed to a subject's head and used to determine position and orientation angles of the head [89]. Since these objective estimates of pose are relatively affordable, they have been the most widely used source of objective ground truth. These products offer a theoretical accuracy of less than 1°, but from our personal experience we have found them to be highly susceptible to noise and the presence of even the smallest amounts of metal in the environment. The environments in which data can be collected are severely restricted, and certain applications such as automotive head pose estimation, are therefore impossible with these sensors.

Inertial Sensors – Inertial sensors utilize accelerometers, gyroscopes, or other motion-sensing devices often coupled with a Kalman filter to reduce noise. The least expensive inertial sensors, such as the Mindflux *InertiaCube*^{2}, do not measure position, but only orientation in 3 DOF. The advantage over magnetic sensors is that there is no metallic interference while attaining similar theoretical accuracy. For head pose estimation, the sensor can be affixed to a subject's head during data capture [81].

Optical Motion Capture Systems – Optical motion capture systems are robust, expensive deployments that are most often used for professional cinematic capture of articulated body movement. Typically, an array of calibrated near-infrared cameras use multiview stereo and software algorithms to follow reflective or active markers attached to a person. For head pose estimation, these markers can be affixed to the back of a subject's head [86] and used to track the absolute position and orientation. Some examples of optical motion capture systems include the Vicon *MX* and the Phoenix Technologies *Visualeyez*.

Using these techniques, a variety of datasets have been collected that range in scope, accuracy, availability, and popularity. Table V on Page 20 contains a description of prominent collections.

B. *Comparison of Published Results*

The mean absolute angular error for pitch, roll, and yaw is a common informative metric for evaluating a head pose estimation system. This metric can be used to evaluate a system on datasets with coarse or fine pose labels, and it provides a single statistic that gives insight into the accuracy of competing methods. Many of the papers that have been discussed here have used this metric for evaluation. Reported results on coarse and fine datasets are described in Table II and Table III respectively. For coarse head pose estimation, it is usual to evaluate an approach by classification error (i.e., how often an image at specific discrete pose angle is correctly labeled with the correct pose label). Although a proof of the validity of the system, the results depend on the number of discrete poses (more discrete poses make for a more challenging dataset). Also, this representation gives little information about the character of each misclassification (was a nearby pose selected, or was the misclassification a widely incorrect estimate?). Regardless of these limitations, classification error has been frequently used to evaluate head pose estimation approaches. Reported results of such errors along with the number of discrete poses in each dataset are described in Table II.

From these tables, a series of observations can be made. On the Pointing '04 dataset, nonlinear regression with MLP neural networks [117] had the lowest reported mean angular error (MAE), demonstrating the powerful representational ability of this nonlinear regression approach to not only estimate head pose, but to learn a mapping that can tolerate the systematic errors in the

TABLE II
MEAN ABSOLUTE ERROR OF COARSE HEAD POSE ESTIMATION

Dataset	**Mean Absolute Error**		**Classification**	**# of Discrete**	
Publication	**Yaw**	**Pitch**	**Accuracy**	**Poses**	**Notes**
Pointing ’04					
J. Wu [137]	-	-	[90%]*	93	-
Stiefelhagen [117]	9.5°	9.7°	{52.0%, 66.3%}†	{13, 9}†	1
Human Performance [34]	11.8°	9.4°	{40.7%, 59.0%}†	{13, 9}†	2
Gourier (Associative Memories) [34]	10.1°	15.9°	{50.0%, 43.9%}†	{13, 9}†	3
Tu (High-order SVD) [125]	12.9°	17.97°	{49.25%, 54.84%}†	{13, 9}†	4
Tu (PCA) [125]	14.11°	14.98°	{55.20%, 57.99%}†	{13, 9}†	4
Tu (LEA) [125]	15.88°	17.44°	{45.16%, 50.61%}†	{13, 9}†	4
Voit [129]	12.3°	12.77°	-	93	-
Zhu (PCA) [66]	26.9°	35.1°	-	93	5
Zhu (LDA) [66]	25.8°	26.9°	-	93	5
Zhu (LPP) [66]	24.7°	22.6°	-	93	5
Zhu (Local-PCA) [66]	24.5°	37.6°	-	93	5
Zhu (Local-LPP) [66]	29.2°	40.2°	-	93	5
Zhu (Local-LDA) [66]	19.1°	30.7°	-	93	5
CHIL-CLEAR06					
Voit [128]	-	-	39.40%	8	-
Voit [129]	49.2°	-	34.90%	8	-
Canton-Ferrer [12]	73.63°	-	19.67%	8	-
Zhang [146]	33.56°	-	[87%]*	5	-
CMU-PIE					
Brown (Probabilistic Method) [10]	3.6°	-	-	9	-
Brown (Neural Network) [10]	4.6°	-	-	9	-
Brown [10]	-	-	91%	9	3
Ba [1]	-	-	71.20%	9	-
Tian [120]	-	-	89%	9	6
Softopia HOIP					
Raytchev (PCA) [101]	15.9°	12.4°	-	91	7
Raytchev (LPP) [101]	15.9°	12.8°	-	91	7
Raytchev (Isomap) [101]	11.5°	11.9°	-	91	8
T. Ma (SVR) [70]	-	-	81.50%	45	34
T. Ma (SBL) [70]	-	-	81.70%	45	34
CVRR-86					
J. Wu (PCA) [136]	-	-	36.40%	86	-
J. Wu (LDA) [136]	-	-	40.10%	86	-
J. Wu (KPCA)[136]	-	-	42%	86	-
J. Wu (KLDA) [136]	-	-	50.30%	86	-
J. Wu (KLDA + EGM) [136]	-	-	75.40%	86	-
CAS-PEAL					
Y. Ma [69]	-	-	97.14%	7	3
Other Datasets					
Niyogi [91]	-	-	48%	15	9
N. Krüger [55]	-	-	92%	5	10
S. Li [63]	-	-	[96.8%]*	10	-
Tian [120]	-	-	84.30%	12	11

Notes

$*$ Any neighboring poses considered correct classification as well
† DOF classified separately {yaw,pitch}

1. Used 80% of Pointing ’04 images for training and 10% for evaluation
2. Human performance with training
3. Best results over different reported methods
4. Better results have been obtained with manual localization
5. Results for 32-dim embedding
6. Best single camera results
7. Results for 100-dim embedding
8. Results for 8-dim embedding
9. Dataset: 15 images for each of 11 subjects
10. Dataset: 413 images of varying people
11. Dataset: Far-field images with wide-baseline stereo

TABLE III

MEAN ABSOLUTE ERROR AND CLASSIFICATION ERROR OF FINE HEAD POSE ESTIMATION

Dataset	**Mean Absolute Error**			
Publication	**Yaw**	**Pitch**	**Roll**	**Notes**
CHIL-CLEAR07				
Canton-Ferrer [13]	20.48°	-	-	-
Voit [130]	8.5°	12.5°	-	-
Yan [142]	6.72°	8.87°	4.03°	
Ba [2]	15°	10°	5°	1
IDIAP Head Pose				
Ba [2]	8.8°	9.4°	9.89°	-
Voit [130]	14.0°	9.2°	-	-
CVRR-363				
Appearance Templates (Cross. Corr.)	21.25°	5.19°	-	-
Replica of Y. Li [64] system [86]	13.46°	5.72°	-	2
Murphy-Chutorian [86]	6.4°	5.58°	-	-
USF HumanID				
Fu [27]	1.7°	-	-	-
BU Face Tracking				
La Cascia [14]	3.3°	6.1°	9.8°	3
Xiao [139]	3.8°	3.2°	1.4°	-
CVRR LISAP-14				
Appearance Templates (Cross. Corr.)	15.69°	7.72°	-	4,5
Replica of Y. Li [64] approach [86]	12.8°	4.66°	-	2,4,5
Murphy-Chutorian [86]	8.25°	4.78°	-	4,5
Murphy-Chutorian [85]	3.39°	4.67°	2.38°	4
FacePix				
Balasubramanian (Isomap)[4]	10.41°	-	-	6
Balasubramanian (LLE)[4]	3.27°	-	-	6
Balasubramanian (LE)[4]	3.93°	-	-	6
Balasubramanian (Biased Isomap)[4]	5.02°	-	-	6
Balasubramanian (Biased LLE)[4]	2.11°	-	-	6
Balasubramanian (Biased LE)[4]	1.44°	-	-	6
Other Datasets				
Y. Li [64]	9.0°	8.6°	-	7,8
Y. Wu [138]	34.6°	14.3°	-	7,9
Sherrah [109]	11.2°	6.4°	-	3,10
L. Zhao [148]	[9.3°]†	[9.0°]†	-	7
Ng [88]	11.2°	8.0°	-	10,11
Stiefelhagen [115]	9.5°	9.1°	-	7,12
Morency [81]	3.5°	2.4°	2.6°	13,14,18
Morency [82]	3.2°	1.6°	1.3°	13,15,18
K. Huang [49]	12°	-	-	16,17,18
Seemann [108]	7.5°	6.7°	-	16,18
Osadchy [95],[96]	12.6°	-	[3.6°]‡	3,19
Hu [46]	6.5°	5.4°	-	20
Oka [94]	2.2°	2.0°	0.7°	15,18
Tu [124]	-	[4°]∗	-	16,21
G. Zhao [147]	[2.44°]⋆	[2.76°]⋆	[2.86°]⋆	22
Wang [132]	1.67	2.56	3.54	13,23

Notes

† Only one DOF allowed to vary at a time
‡ In-plane rotation instead of roll
⋆ Root Mean Square (RMS) error
∗ Error combined over pitch, roll, and yaw

1. Best results over different reported methods
2. Results for 50-dim embedding
3. Error estimated from plots
4. Averaged results for monocular nighttime and daytime sequences
5. Dataset: Uniformly sampled 5558 pose images from driving sequences
6. Results for 100-dim embedding
7. Data collected with magnetic sensor
8. Dataset: 1283 images covering 12 people
9. Dataset: 16 video sequences covering 11 people
10. Dataset: Two 200-frame video sequences
11. Results for test subjects not included in training set
12. Dataset: 760 images covering 2 people
13. Data collected with an inertial sensor
14. Dataset: 4 video sequences
15. Dataset: 2 video sequences
16. Dataset: 1 video sequence
17. Best results over different reported methods
18. Uses stereo imagery
19. Dataset: 388 images
20. Dataset: 5 sequences, each with 20 views and limited pitch variation
21. Error averaged over results from four resolutions
22. Dataset: 3 video sequences
23. Dataset: 400 frames for each of 8 subjects

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

training data, discussed in Table V on page 20. In comparison, people do not learn invariance to this error and demonstrate notably worse performance when asked to perform the similar task of yaw estimation [34]. On the multi-camera CHIL-CLEAR07 evaluative dataset, manifold embedding with SSE provided the most accurate results. This suggests that manifold embedding approaches can provide superior representational ability, although of all the techniques presented in this paper, only the linear SSE embedding has been evaluated with inhomogeneous training data.

A series of comparisons have been made for different manifold embedding approaches. On the Pointing '04 dataset, Local-LDA produced better yaw estimation than PCA, LDA, LPP, Local-PCA, and Local-LPP [66], but for pitch estimation, however, standard LDA provided better results than these other techniques. The fact that the localized versions of these embeddings do not uniformly improve pose estimation may be limited by the ability to choose the correct local projection for each new sample. On the CVRR-86 dataset, KLDA was shown to outperform PCA, LDA, and KPCA, clearly demonstrating that the kernelized versions provide a better embedding for pose estimation [136]. On the Softopia HOIP dataset it has been shown that a projection into 8-dimensions with Isomap is enough to obtain superior results over PCA and LPP subspaces with 100 dimensions [101]. This should motivate continued investigation of nonlinear embedding methods as the increase in representational ability can lead to large improvements in pose estimation.

The methods that track the head in video sequences using flexible models and tracking methods [14, 81, 82, 94, 124, 140, 147] report a significantly lower error than systems that estimate pose in individual images. Although these systems use different datasets that cannot be directly compared, from our experience, visual tracking approaches provide substantially less error than systems that estimate the head pose error from individual video frames and temporally filter the results.

C. Comparison of Real-World Applicability

For a head pose estimation system to be of general use, it should be invariant to identity, have sufficient range of allowed motion, require no manual intervention, and should be easily deployed on conventional hardware. Although some systems address all of these concerns, they often assume one or more conditions that simplify the pose estimation problem, at the expense of general applicability. We have identified the following set of assumptions that have been commonly used in the literature:

A. **Continuous video assumption**. There is a continuous video stream with only small pose changes between frames, so that head pose can be estimated by incremental relative shifts between each frame.

B. **Initialization assumption**. The head pose of the subject is known when the pose estimation process begins. In practice, the subject is typically told to assume a frontal pose until the system has begun, or the system waits until a frontal facial detector discovers a frontal pose.

C. **Anti-Drift assumption**. Head pose will only be calculated for a short time, during which there will be no significant abnormalities from the visual information. If this assumption were violated, the pose estimation system would be subject to drift, and continued estimates of pose will have large errors.

D. **Stereo vision assumption**. Subject is visible by two or more cameras at a small enough distance to discriminate depth information across the face. Alternatively, depth information can be obtained by other specialized means, such as a time-of-flight sensor [150].

E. **Frontal view assumption**. Pose variation is restricted to a range that contains all of the facial features seen from a frontal view.

F. **Single degree of freedom assumption**. The head is only allowed to rotate around one axis.

G. **Feature location assumption**. The location of facial features are provided to the system. This typically implies that facial features are manually labeled in the testing data.

H. **Familiar-identity assumption**. The system needs to estimate pose only for the person or set of persons on which it has been trained.

I. **Synthetic data assumption**. The system only operates on synthetic images that do not contain the appearance variation found in real-world imagery.

These assumptions limit the applicability of any system, even when shown to be quite successful in a constrained environment. Regardless of estimation accuracy, it is important to identify the systems which are applicable for head pose estimation in real world environments such as the automobile and intelligent spaces. Such systems should provide an identity-invariant estimate of head pose in at least two degrees-of-freedom without any manual intervention. For systems with discrete pose estimates, the number of fixed poses must be large enough to sufficiently sample the continuous pose space. These systems are indicated in bold type in Table IV, which contains a comprehensive list of all the papers covered in this survey.

V. Summary and Concluding Remarks

Head pose estimation is a natural step for bridging the information gap between people and computers. This fundamental human ability provides rich information about the intent, motivation, and attention of people in the world. By simulating this skill, systems can be created that can better interact with people. The majority of head pose estimation approaches assume the perspective of a rigid model, which has inherent limitations. The difficulty in creating head pose estimation systems stems from the immense variability in individual appearance coupled with differences in lighting, background, and camera geometry.

Viewing the progress in head pose estimation chronologically, we have noticed some encouraging progress in the field. In recent years, people have become much more aware of the need for comparison metrics that emphasize pose variation rather than image variation. This trend has manifested itself with the demise of appearance templates and the explosion of nonlinear manifold embedding methods. Coarse head pose estimation is also disappearing as most recent work focuses on fine estimation and multiple degrees-of-freedom. As a result, new datasets have been introduced in the past few years that allow for more accurate evaluation in challenging environments. We feel there is still much room for continued improvement. Model-based tracking algorithms have shown great promise, but they will require more thorough evaluations on standard datasets to understand their potential. Geometric methods have not reached their full potential, yet modern approaches can automatically and reliably detect facial feature locations, and these approaches should continue

TABLE IV

HEAD POSE ESTIMATION PUBLICATIONS.

Year	Publication	Approach	Assump.	DOF	Domain	View	Automatic
1994	Gee [30]	Geometric	G	2	continuous	near	no
1994	Beymer [7]	Appearance Templ.	E	2	15 poses	near	yes
1995	Lanitis [60, 61]	Manifold Embed.	E	2	continuous	near	no
1995	Schiele [106]	Nonlinear Regress.	F	1	15 poses	near	yes
1996	**Maurer [73]**	**Tracking**	**A,B,C**	**3**	**continuous**	**near**	**yes**
1996	Horprasert [42]	Geometric	G	3	continuous	near	no
1996	Niyogi [91]	Appearance Templ.	-	2	15 poses	near	no
1996	Gee [31]	Tracking	A,B,C,G	3	continuous	near	no
1997	Horprasert [43]	Hybrid	A,B,C,G	3	continuous	near	no
1997	**Jebara [52]**	**Hybrid**	**A,C**	**3**	**continuous**	**near**	**yes**
1997	N. Krüger [55]	Flexible Models	E,F	1	5 poses	near	yes
1998	Rowley [103]	Detector Arrays	F	1	continuous	far	yes
1998	**Rae [100]**	**Nonlinear Regress.**	**-**	**2**	**35 poses**	**near**	**yes**
1998	**Schödl [107]**	**Tracking**	**A,B**	**3**	**continuous**	**near**	**yes**
1998	Bruske [11]	Nonlinear Regress.	H,I	2	continuous	near	no
1998	McKenna [75]	Manifold Embed.	H	2	49 poses	near	yes
1998	**Pappu [98]**	**Tracking**	**-**	**2**	**187 poses**	**near**	**yes**
1998	**Toyama [121]**	**Tracking**	**-**	**3**	**continuous**	**near**	**yes**
1998	J. Huang [47]	Detector Arrays	E,F	1	3 poses	near	yes
1999	**Ng [88, 89]**	**Appearance Templ.**	**-**	**2**	**continuous**	**near**	**yes**
1999	Sherrah [110, 111]	Manifold Embed.	-	1†	19 or 7 poses	near	no
2000	**DeCarlo [21]**	**Tracking**	**A,B,G**	**3**	**continuous**	**near**	**yes**
2000	**Y. Li [64, 65]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**near**	**yes**
2000	Malciu [71]	Tracking	A,B,C	3	continuous	near	no
2000	Cootes [18]	Flexible Models	F	1	continuous	near	no
2000	**Y. Wu [138]**	**Tracking**	**-**	**3**	**1024 poses**	**far**	**yes**
2000	**La Cascia [14]**	**Tracking**	**A,B,C**	**3**	**continuous**	**near**	**yes**
2000	Nikolaidis [90]	Geometric	E,F	1	continuous	near	yes
2001	Cordea [20]	Geometric	F	1	continuous	near	yes
2001	**Yao [145]**	**Tracking**	**A,B,C,G**	**2**	**continuous**	**near**	**yes**
2001	S. Li [63]	Detector Arrays	F	1	10 poses	far	yes
2001	**Sherrah [109]**	**Hybrid**	**A,B**	**2**	**continuous**	**near**	**yes**
2002	V. Krüger [56]	Manifold Embed.	E,H,I	2	continuous	near	no
2002	L. Zhao [148]	Nonlinear Regress.	F	1†	19 poses	near	no
2002	**Morency [80]**	**Tracking**	**A,B,C,D**	**3**	**continuous**	**near**	**yes**
2002	Brown [10]	Nonlinear Regress.	F	1	9 poses	near	no
2002	Yang [144]	Tracking	A,B,C,D,G	3	continuous	near	no
2002	Srinivasan [114]	Manifold Embed.	F	1	7 poses	near	no
2002	**Stiefelhagen [115, 117]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**both**	**yes**
2003	**Morency [81, 82]**	**Hybrid**	**A,B,D**	**3**	**continuous**	**near**	**yes**
2003	**Zhu [150]**	**Tracking**	**A,B,C,D**	**3**	**continuous**	**near**	**yes**
2003	**Xiao [139]**	**Tracking**	**A,B,C**	**3**	**continuous**	**near**	**yes**
2003	Chen [15]	Detector Arrays	E,F,H,I	1	continuous	near	yes
2003	Tian [120]	Nonlinear Regress.	D,F	1	up to 12 poses	far	yes
2003	Jones [53]	Detector Arrays	F	1	12 poses	far	yes
2004	**Osadchy [95, 96]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**far**	**yes**
2004	Raytchev [101]	Manifold Embed.	-	2	continuous	near	no
2004	K. Huang [49]	Hybrid	A,F	1	38 poses	near	yes
2004	**Seemann [108]**	**Nonlinear Regress.**	**D**	**2**	**continuous**	**far**	**yes**
2004	**Ba [1, 2]**	**Hybrid**	**-**	**3**	**continuous**	**far**	**yes**
2004	**Gourier [33, 34]**	**Nonlinear Regress.**	**-**	**2**	**13 yaws, 9 pitches**	**near**	**yes**
2004	**Xiao [140]**	**Flexible Models**	**E**	**3**	**continuous**	**near**	**yes**
2004	**Y. Hu [46]**	**Hybrid**	**A,C**	**2**	**continuous**	**near**	**yes**
2004	**Dornaika [22]**	**Tracking**	**A,B**	**3**	**continuous**	**near**	**yes**
2004	**Baker [3]**	**Flexible Models**	**E**	**3**	**continuous**	**near**	**yes**
2004	Moon [78]	Nonlinear Regress.	G	2	continuous	near	no
2004	**Zhu [151]**	**Hybrid**	**A**	**3**	**continuous**	**near**	**yes**
2004	**C. Hu [44]**	**Flexible Models**	**E,D**	**3**	**continuous**	**near**	**yes**
2005	**Oka [94]**	**Tracking**	**A,B,D,C**	**3**	**continuous**	**near**	**yes**
2005	**Xiong [141]**	**Geometric**	**A,D**	**2**	**continuous**	**near**	**yes**
2005	N. Hu [45]	Manifold Embed.	-	1	continuous	near	no
2005	**J. Wu [136, 137]**	**Hybrid**	**-**	**2**	**86 poses**	**near**	**yes**
2006	Fu [27]	Manifold Embed.	F	1	continuous	near	no
2006	**Voit [128–130]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**both**	**yes**
2006	B. Ma [69]	Detector Arrays	-	1	7 poses	near	yes
2006	**Y. Ma [70]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**near**	**yes**
2006	Ohayon [93]	Tracking	A,G	3	continuous	near	no
2006	**Gui [35]**	**Flexible Models**	**E,A**	**3**	**continuous**	**near**	**yes**
2006	**Tu [124]**	**Tracking**	**A,B,C**	**3**	**continuous**	**far**	**yes**
2006	Canton-Ferrer [12, 13]	Geometric	F,D	1	continuous	far	yes
2006	Tu [125]	Tracking	-	2	13 yaws, 9 pitches	near	no
2006	Zhang [146]	Detector Arrays	F	1	5 poses	far	yes
2007	**Murphy-Chutorian [86]**	**Nonlinear Regress.**	**-**	**2**	**continuous**	**near**	**yes**
2007	Yan [142]	Manifold Embed.	-	2	continuous	far	no
2007	Balasubramanian [4]	Manifold Embed.	F	1	continuous	near	no
2007	**G. Zhao [147]**	**Tracking**	**A,D**	**3**	**continuous**	**far**	**yes**
2007	Zhu [66]	Manifold Embed.	-	2	93 poses	near	no
2007	**Wang [132]**	**Geometric**	**E,G**	**3**	**continuous**	**near**	**yes**
2008	**Murphy-Chutorian [85]**	**Hybrid**	**A**	**3**	**continuous**	**near**	**yes**

(Systems in bold face provide fully automatic, fine, identity-invariant head pose estimation in at least 2 DOF)

†Only one degree of freedom allowed to vary at a time.

to evolve. Another important trend observed is an increase in the number of head pose publications over the last few years. This might be a sign that more people have become interested in this field, suggesting a more rapid developmental cycle for novel approaches.

Although head pose estimation will continue to be an exciting field with much room for improvement, people desire off-the-shelf, universal head pose estimation programs that can be put to use in any new application. To satisfy the majority of applications, we propose the following design criteria as a guide for future development.

- **Accurate**: The system should provide a reasonable estimate of pose with a mean absolute error of 5° or less.
- **Monocular**: The system should be able to estimate head pose from a single camera. Although accuracy might be improved by stereo or multi-view imagery, this should not be a requirement for the system to operate.
- **Autonomous**: There should be no expectation of manual initialization, detection, or localization, precluding the use of pure-tracking approaches that measure the relative head pose with respect to some initial configuration and shape/geometric approaches that assume facial feature locations are already known.
- **Multi-Person**: The system should be able to estimate the pose of multiple people in one image.
- **Identity & Lighting Invariant**: The system must work across all identities with the dynamic lighting found in many environments.
- **Resolution Independent**: The system should apply to near-field and far-field images with both high and low resolution.
- **Full Range of Head Motion**: The methods should be able to provide a smooth, continuous estimate of pitch, yaw, and roll, even when the face is pointed away from the camera.
- **Real-time**: The system should be able to estimate a continuous range of head orientation with fast (30fps or faster) operation.

Although no single system has met all of these criteria, it does seem that a solution is near at hand. It is our opinion that by using today's methods in a suitable hybrid approach (perhaps a combination of manifold embedding, regression, or geometric methods combined with a model-based tracking system), one could meet these criteria.

For future work, we would expect to see an evaluation of nonlinear manifold embedding techniques on challenging far-field imagery, demonstrating that these approaches provide continued improvement in the presence of clutter or imperfect localization. We would like to see extensions for geometric and tracking approaches that adapt the models to each subject's personal facial geometry for more accurate model fitting. For flexible models, an important improvement will be the ability to selectively ignore parts of the model are self-occluded, overcoming a fundamental limitation in an otherwise very promising category. In closing, we describe a few application domains in which head pose estimation has and will continue to have a profound impact.

Head pose estimation systems will play a key role in the creation of intelligent environments. Already there has been a huge interest in smart rooms that monitor the occupants and use head pose to measure their activities and visual focus of attention [5, 50, 74, 92, 99, 116, 122–124, 128, 131]. Head pose endows these systems with the ability to determine who is speaking to whom, and to provide the information needed to analyze the non-verbal gestures of the meeting participants. These type of high-level semantic cues can be transcribed along with the conversations, intentions, and interpersonal interactions of the meeting participants to provide easily searchable indexes for future reference.

Head pose estimation could enable breakthrough interfaces for computing. Some existing examples include systems that allow a user to control a computer mouse using his head pose movements [28], respond to pop-up dialog boxes with head nods and shakes [84], or use head gestures to interact with embodied agents [83]. It seems only a matter of time until similar estimation algorithms are integrated into entertainment devices with mass-appeal.

Head pose estimation will have a profound impact on the future of automotive safety. An automobile driver is fundamentally limited by the field-of-view that one can observe at any one time. When one fails to notice a change to his environment, there is an increased potential for a life-threatening collision that could be mitigated if the driver was alerted to an unseen hazard. As evidence to this effect, a recent comprehensive survey on automotive collisions demonstrated a driver was 31% less likely to cause an injury-related collision when there was one or more passengers [105]. Consequently, there is great interest in driver assistance systems that act as virtual passengers, using the driver's head pose as a cue to visual focus of attention and mental state [3, 16, 36, 48, 85, 86, 98, 135, 151]. Although the rapidly shifting lighting conditions in a vehicle make for one of the most difficult visual environments, the most recent of these systems demonstrated a fully-automatic, real-time hybrid approach that can estimate the pose and track the driver's head in daytime or nighttime driving [85].

We believe that ubiquitous head pose estimation is just beyond the grasp of our current systems, and we call on researchers to improve and extend the techniques described in this paper to allow life-changing advances in systems for human interaction and safety.

Acknowledgment

Support for this research was provided by the University of California Discovery Program and the Volkswagen Electronics Research Laboratory. We sincerely thank our reviewers for their constructive and insightful suggestions that have helped improve the quality of this paper. In addition, we thank our colleagues in the Computer Vision and Robotics Research laboratory for their valuable assistance.

References

[1] S. Ba and J.-M. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *Proc. Int'l. Conf. Pattern Recognition*, 2004, pp. 264–267.

[2] ——, "From camera head pose to 3d global room head pose using multiple camera views," in *Proc. Int'l. Workshop Classification of Events Activities and Relationships*, 2007.

[3] S. Baker, I. Matthews, J. Xiao, R. Gross, T. Kanade, and T. Ishikawa, "Real-time non-rigid driver head tracking for driver mental state estimation," in *Proc. 11th World Congress Intelligent Transportation Systems*, 2004.

[4] V. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent

head pose estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[5] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Towards measuring human interactions in conversational settings," in *Proc. IEEE Int'l Workshop Cues in Communication*, 2001.

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[7] D. Beymer, "Face recognition under varying pose," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 756–761.

[8] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[9] K. Bowyer and S. Sarkar, "USF HumanID 3D face dataset," 2001.

[10] L. Brown and Y.-L. Tian, "Comparative study of coarse head pose estimation," in *Proc. Workshop Motion and Video Computing*, 2002, pp. 125–130.

[11] J. Bruske, E. Abraham-Mumm, J. Pauli, and G. Sommer, "Head-pose estimation from facial images with subspace neural networks," in *Proc. Int'l. Neural Network and Brain Conference*, 1998, pp. 528–531.

[12] C. Canton-Ferrer, J. Casas, and M. Pardàs, "Head pose detection based on fusion of multiple viewpoint information," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 305–310.

[13] ——, "Head orientation estimation using particle filtering in multiview scenarios," in *Proc. Int'l. Workshop Classification of Events Activities and Relationships*, 2007.

[14] M. L. Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 4, pp. 322–336, 2000.

[15] I. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang, "Head pose estimation using Fisher Manifold learning," in *Proc. IEEE Int'l. Workshop Analysis and Modeling of Faces and Gestures*, 2003, pp. 203–207.

[16] S. Cheng, S. Park, and M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, 2006.

[17] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models–their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[18] T. Cootes, K. Walker, and C. Taylor, "View-based active appearance models," in *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 227–232.

[19] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[20] M. Cordea, E. Petriu, N. Georganas, D. Petriu, and T. Whalen, "Real-time 2(1/2)-D head pose recovery for model-based video-coding," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 4, pp. 1007–1013, 2001.

[21] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Int'l. J. Computer Vision*, vol. 38, no. 2, pp. 231–238, 2000.

[22] F. Dornaika and F. Davoine, "Head and facial animation tracking using appearance-adaptive models and particle filters," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop*, 2004, pp. 153–162.

[23] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, Inc., 2001.

[24] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. F. Cootes, "Statistical models of face images - improving specificity," *Image Vision Computing*, vol. 16, no. 3, pp. 203–211, 1998.

[25] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[26] V. F. Ferrario, C. Sforza, G. Serrao, G. Grassi, and E. Mossi, "Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults," *J. Orthopaedic Research*, vol. 20, no. 1, pp. 122–129, 2002.

[27] Y. Fu and T. Huang, "Graph embedded analysis for head pose estimation," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2006, pp. 3–8.

[28] Y. Fu and T. S. Huang, "hMouse: Head tracking driven virtual computer mouse," in *IEEE Workshop Applications of Computer Vision*, 2007, pp. 30–35.

[29] W. Gao, B. Cao, S. Shan, X. Zhang, and D. Zhou, "The CAS-PEAL large-scale chinese face database and baseline evaluations," Joint Research & Development Laboratory, Tech. Rep., 2004, JDL-TR-04-FR-001.

[30] A. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, no. 10, pp. 639–647, 1994.

[31] ——, "Fast visual tracking by temporal consensus," *Image and Vision Computing*, vol. 14, no. 2, pp. 105–114, 1996.

[32] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed. Prentice-Hall, Inc., 2002, pp. 582–584.

[33] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, 2004, pp. 17–25.

[34] N. Gourier, J. Maisonnasse, D. Hall, and J. Crowley, "Head pose estimation on low resolution images," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 270–280.

[35] Z. Gui and C. Zhang, "3D head pose estimation using non-rigid structure-from-motion and point correspondence," in *Proc. IEEE Region 10 Conf.*, 2006, pp. 1–3.

[36] Z. Guo, H. Liu, Q. Wang, and J. Yang, "A fast algorithm face detection and head pose estimation for driver assistant system," in *Proc. Int'l Conf. Signal Processing*, vol. 3, 2006.

[37] M. Harville, A. Rahimi, T. Darrell, G. Gordon, and J. Woodfill, "3D pose tracking with linear depth and brightness constraints," in *Proc. IEEE Int'l. Conf. Computer Vision*, 1999, pp. 206–213.

[38] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice-Hall, Inc., 2002.

[39] X. He, S. Yan, Y. Hu, and H. J. Zhang, "Learning a locality preserving subspace for visual recognition," in *Proc. IEEE Int'l. Conf. Computer Vision*, 2003, pp. 385–392.

[40] J. Heinzmann and A. Zelinsky, "3-D facial pose and gaze point estimation using a robust real-time tracking paradigm," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 142–147.

[41] E. Hjelmås and B. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[42] T. Horprasert, Y. Yacoob, and L. Davis, "Computing 3-d head orientation from a monocular image sequence," in *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 242–247.

[43] ——, "An anthropometric shape model for estimating head orientation," in *Proc. Int'l. Workshop Visual Form*, 1997, pp. 247–256.

[44] C. Hu, J. Xiao, I. Matthews, S. Baker, J. Cohn, and T. Kanade, "Fitting a single active appearance model simultaneously to multiple images," in *Proc. British Machine Vision Conference*, 2004, pp. 437–446.

[45] N. Hu, W. Huang, and S. Ranganath, "Head pose estimation by non-linear embedding and mapping," in *Proc. IEEE Int'l. Conf. Image Processing*, vol. 2, 2005, pp. 342–345.

[46] Y. Hu, L. Chen, Y. Zhou, and H. Zhang, "Estimating face pose by facial asymmetry and geometry," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2004, pp. 651–656.

[47] J. Huang, X. Shao, and H. Wechsler, "Face pose discrimination using support vector machines (SVM)," in *Proc. Int'l. Conf. Pattern Recognition*, 1998, pp. 154–156.

[48] K. Huang and M. Trivedi, "Video arrays for real-time tracking of person, head, and face in an intelligent room," *Machine Vision and Applications*, vol. 14, no. 2, pp. 103–111, 2003.

[49] ——, "Robust real-time detection, tracking, and pose estimation of faces in video streams," in *Proc. Int'l. Conf. Pattern Recognition*, 2004, pp. 965–968.

[50] K. Huang, M. Trivedi, and T. Gandhi, "Driver's view and vehicle surround estimation using omnidirectional video stream," in *Proc. IEEE Intelligent Vehicles Symposium*, 2003, pp. 444–449.

[51] M. Isard and A. Blake, "CONDENSATION–conditional density propagation for visual tracking," *Int'l. J. Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[52] T. Jebara and A. Pentland, "Parametrized structure from motion for 3d adaptive feedback tracking of faces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 144–150.

[53] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Electric Research Laboratories, Tech. Rep. 096, 2003.

[54] H. Kobayasi and S. Kohshima, "Unique morphology of the human eye," *Nature*, vol. 387, no. 6635, pp. 767–768, 1997.

[55] N. Krüger, M. Pötzsch, and C. von der Malsburg, "Determination of face position and pose with a learned representation based on labeled graphs," *Image and Vision Computing*, vol. 15, no. 8, pp. 665–673, 1997.

[56] V. Krüger and G. Sommer, "Gabor wavelet networks for efficient head pose estimation," *Image and Vision Computing*, vol. 20, no. 9-10, pp. 665–672, 2002.

[57] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, pp. 300–311, 1993.

[58] S. Langton and V. Bruce, "You must see the point: Automatic processing of cues to the direction of social attention," *J. Experimental Psychology: Human Perception and Performance*, vol. 26, no. 2, pp. 747–757, 2000.

[59] S. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception and Psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

[60] A. Lanitis, C. Taylor, and T. Cootes, "Automatic interpretation of human faces and hand gestures using flexible models," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1995, pp. 98–103.

[61] ——, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 743–756, 1997.

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[63] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *Proc. IEEE Int'l. Conf. Computer Vision*, 2001, pp. 674–679.

[64] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 300–305.

[65] Y. Li, S. Gong, J. Sherrah, and H. Liddell, "Support vector machine based multi-view face detection and recognition," *Image and Vision Computing*, vol. 22, no. 5, p. 2004, 2004.

[66] Z. Li, Y. Fu, J. Yuan, T. Huang, and Y. Wu, "Query driven localized linear discriminant models for head pose estimation," in *Proc. IEEE Int'l. Conf. Multimedia and Expo*, 2007, pp. 1810–1813.

[67] D. Little, S. Krishna, J. Black, and S. Panchanathan, "A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle," in *Proc. IEEE Int'l. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, 2005, pp. 89–92.

[68] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[69] B. Ma, W. Zhang, S. Shan, X. Chen, and W. Gao, "Robust head pose estimation using LGBP," in *Proc. IEEE Int'l. Conf. Pattern Recognition*, 2006, pp. 512–515.

[70] Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade, "Sparse bayesian regression for head pose estimation," in *Proc. IEEE Int'l. Conf. Pattern Recognition*, 2006, pp. 507–510.

[71] M. Malciu and F. Preteux, "A robust model-based approach for 3D head tracking in video sequences," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 169–174.

[72] I. Matthews and S. Baker, "Active appearance models revisited," *Int'l. J. Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[73] T. Maurer and C. von der Malsburg, "Tracking and learning graphs and pose on image sequences of faces," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 176–181.

[74] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, 2005.

[75] S. McKenna and S. Gong, "Real-time face pose estimation," *Real-Time Imaging*, vol. 4, no. 5, pp. 333–347, 1998.

[76] T. Moeslund, A. Hilton, and V. Krüger, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[77] ——, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[78] H. Moon and M. Miller, "Estimating facial pose from a sparse representation," in *Proc. Int'l. Conf. Image Processing*, 2004, pp. 75–78.

[79] M. Morales, P. Mundy, C. Delgado, M. Yale, R. Neal, and H. Schwartz, "Gaze following, temperament, and language development in 6-month-olds: A replica and extension," *Infant Behavior and Development*, vol. 23, no. 2, pp. 231–236, 2000.

[80] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell, "Fast stereo-based head tracking for interactive environments," in *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 375–380.

[81] L.-P. Morency, A. Rahimi, and T. Darrell, "Adaptive view-based appearance models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 803–810.

[82] L.-P. Morency, P. Sundberg, and T. Darrell, "Pose estimation using 3d view-based eigenspaces," in *Proc. IEEE Int'l. Workshop Analysis and Modeling of Faces and Gestures*, 2003, pp. 45–52.

[83] L.-P. Morency, C. Christoudias, and T. Darrell, "Recognizing gaze aversion gestures in embodied conversational discourse," in *Proc. Int'l. Conf. Multimodal Interfaces*, 2006, pp. 287–294.

[84] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artificial Intelligence*, vol. 171, no. 8–9, pp. 568–585, 2007.

[85] E. Murphy-Chutorian and M. Trivedi, "Hybrid head orientation

and position estimation (HyHOPE): A system and evaluation for driver support," in *Proc. IEEE Intelligent Vehicles Symposium*, 2008.

[86] ——, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *Proc. IEEE Conf. Intelligent Transportation Systems*, 2007, pp. 709–714.

[87] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky, "Real-time stereo tracking for head pose and gaze estimation," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 122–128.

[88] J. Ng and S. Gong, "Composite support vector machines for detection of faces across views and pose estimation," *Image and Vision Computing*, vol. 20, no. 5-6, pp. 359–368, 2002.

[89] ——, "Multi-view face detection and pose estimation using a composite support vector machine across the view sphere," in *Proc. IEEE Int'l. Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 1999, pp. 14–21.

[90] A. Nikolaidis and I. Pitas, "Facial feature extraction and pose determination," *Pattern Recognition*, vol. 33, no. 11, pp. 1783–1791, 2000.

[91] S. Niyogi and W. Freeman, "Example-based head tracking," in *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 374–378.

[92] J.-M. Odobez and S. Ba, "A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose," in *Proc. IEEE Int'l. Conf. Multimedia and Expo*, 2007, pp. 1379–1382.

[93] S. Ohayon and E. Rivlin, "Robust 3D head tracking using camera pose estimation," in *Proc. IEEE Int'l. Conf. Pattern Recognition*, 2006, pp. 1063–1066.

[94] K. Oka, Y. Sato, Y. Nakanishi, and H. Koike, "Head pose estimation system based on particle filtering with adaptive diffusion control," in *Proc. IAPR Conf. Machine Vision Applications*, 2005, pp. 586–589.

[95] R. Osadchy, M. Miller, and Y. LeCun, "Synergistic face detection and pose estimation with energy-based model," in *Proc. Advances in Neural Information Processing Systems*, 2004, pp. 1017–1024.

[96] ——, "Synergistic face detection and pose estimation with energy-based models," *J. Machine Learning Research*, vol. 8, pp. 1197–1215, 2007.

[97] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 130–136.

[98] R. Pappu and P. Beardsley, "A qualitative approach to classifying gaze direction," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 160–165.

[99] A. Pentland and T. Choudhury, "Face recognition for smart environments," *Computer*, vol. 33, no. 2, pp. 50–55, 2000.

[100] R. Rae and H. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Trans. Neural Networks*, vol. 9, no. 2, pp. 257–265, 1998.

[101] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in *Proc. Int'l. Conf. Pattern Recognition*, 2004, pp. 462–466.

[102] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[103] H. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 38–44.

[104] ——, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 1, pp. 23–38, 1998.

[105] T. Rueda-Domingo, P. Lardelli-Claret, J. L. del Castillo, J. Jiménez-Moleón, M. García-Martín, and A. Bueno-Cavanillas, "The influence of passengers on the risk of the driver causing a car collision in spain," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 481–489, 2004.

[106] B. Schiele and A. Waibel, "Gaze tracking based on face-color," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1995, pp. 344–349.

[107] A. Schödl, A. Haro, and I. Essa, "Head tracking using a textured polygonal model," in *Proc. Workshop Perceptual User Interfaces*, 1998.

[108] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2004, pp. 626–631.

[109] J. Sherrah and S. Gong, "Fusion of perceptual cues for robust tracking of head pose and position," *Pattern Recognition*, vol. 34, no. 8, pp. 1565–1572, 2001.

[110] J. Sherrah, S. Gong, and E.-J. Ong, "Understanding pose discrimination in similarity space," in *British Machine Vision Conference*, 1999, pp. 523–532.

[111] ——, "Face distributions in similarity space under varying head pose," *Image and Vision Computing*, vol. 19, no. 12, pp. 807–819, 2001.

[112] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[113] Softopia, "HOIP face database." [Online]. Available: http://www.softopia.or.jp/en/rd/facedb.html

[114] S. Srinivasan and K. Boyer, "Head pose estimation using view based eigenspaces," in *Proc. Int'l. Conf. Pattern Recognition*, 2002, pp. 302–305.

[115] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[116] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proc. Int'l. Conf. Multimodal Interfaces*, 2002, pp. 273–280.

[117] ——, "Estimating head pose with neural networks - results on the Pointing04 ICPR workshop evaluation data," in *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, 2004.

[118] R. Stiefelhagen, K. Bernardin, R. B. J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 1–44.

[119] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[120] 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 *Proc. IEEE Int'l. Workshop Analysis and Modeling of Faces and Gestures*, 2003, pp. 92–99.

[121] K. Toyama, ""look, ma – no hands!" hands-free cursor control with real-time 3d face tracking," in *Proc. Workshop Perceptual User Interfaces*, 1998, pp. 49–54.

[122] M. Trivedi, "Human movement capture and analysis in intelligent environments," *Machine Vision and Applications*, vol. 14, no. 4, pp. 215–217, 2003.

[123] M. Trivedi, K. Huang, and I. Mikić, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Trans. Syst., Man, Cybern. A*, vol. 35, no. 1, pp. 145–163, 2005.

[124] J. Tu, T. Huang, and H. Tao, "Accurate head pose tracking in low resolution video," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2006, pp. 573–578.

[125] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds.,

vol. 4122, 2007, pp. 281–290.

[126] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[127] M. Voit, "CLEAR 2007 evaluation plan: Head pose estimation," 2007. [Online]. Available: http://isl.ira.uka.de/~mvoit/clear07/CLEAR07_HEADPOSE_2007-03-26.doc

[128] M. Voit, K. Nickel, and R. Stiefelhagen, "A bayesian approach for multi-view head pose estimation," in *IEEE Int'l. Conf. Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 31–34.

[129] ——, "Neural network-based head pose estimation and multi-view fusion," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 291–298.

[130] ——, "Head pose estimation in single- and multi-view environments results on the CLEAR'07 benchmarks," in *Proc. Int'l. Workshop Classification of Events Activities and Relationships*, 2007.

[131] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "SMaRT: the smart meeting room task at isl," in *IEEE Int'l. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, 2003, pp. 752–755.

[132] J.-G. Wang and E. Sung, "EM enhancement of 3D head pose estimated by point at infinity," *Image and Vision Computing*, vol. 25, no. 12, pp. 1864–1874, 2007.

[133] H. Wilson, F. Wilkinson, L. Lin, and M. Castillo, "Perception of head orientation," *Vision Research*, vol. 40, no. 5, pp. 459–472, 2000.

[134] W. H. Wollaston, "On the apparent direction of eyes in a portrait," *Phil. Trans. Royal Society of London*, vol. 114, pp. 247–256, 1824.

[135] J.-W. Wu and M. Trivedi, "Visual modules for head gesture analysis in intelligent vehicle systems," in *Proc. IEEE Intelligent Vehicles Symposium*, 2006, pp. 13–18.

[136] J. Wu and M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recognition*, vol. 41, no. 3, pp. 1138–1158, 2008.

[137] J. Wu, J. Pedersen, D. Putthividhya, D. Norgaard, and M. M. Trivedi, "A two-level pose estimation framework using majority voting of Gabor wavelets and bunch graph analysis," in *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, 2004.

[138] Y. Wu and K. Toyama, "Wide-range, person- and illumination-insensitive head orientation estimation," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 183–188.

[139] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn, "Robust full-motion recovery of head by dynamic templates and re-registration techniques," *Int'l. J. Imaging Systems and Technology*, vol. 13, no. 1, pp. 85–94, 2003.

[140] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 535–542.

[141] Y. Xiong and F. Quek, "Meeting room configuration and multiple cameras calibration in meeting analysis," in *Int'l. Conf. Multimodal Interfaces*, 2005, pp. 1–8.

[142] S. Yan, Z. Zhang, Y. Fu, Y. Hu, J. Tu, , and T. Huang, "Learning a person-independent representation for precise 3d pose estimation," in *Proc. Int'l. Workshop Classification of Events Activities and Relationships*, 2007.

[143] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 1, pp. 34–58, 2002.

[144] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 242–247.

[145] P. Yao, G. Evans, and A. Calway, "Using affine correspondence to estimate 3-d facial pose," in *Proc. Int'l. Conf. Image Processing*, 2001, pp. 919–922.

[146] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," in *Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 299–304.

[147] G. Zhao, L. Chen, J. Song, and G. Chen, "Large head movement tracking using SIFT-based registration," in *Proc. Int'l. Conf. Multimedia*, 2007, pp. 807–810.

[148] L. Zhao, G. Pingali, and I. Carlbom, "Real-time head orientation estimation using neural networks," in *Proc. Int'l. Conf. Image Processing*, 2002, pp. 297–300.

[149] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[150] Y. Zhu and K. Fujimura, "3D head pose estimation with optical flow and depth constraints," in *Proc. IEEE Int'l. Conf. 3-D Digital Imaging and Modeling*, 2003, pp. 211–216.

[151] ——, "Head pose estimation for driver monitoring," in *Proc. IEEE Intelligent Vehicles Symposium*, 2004, pp. 501–506.

Erik Murphy-Chutorian received the B.A. degree in Engineering Physics from Dartmouth College in 2002. He received the M.S. degree in Electrical and Computer Engineering from the University of California, San Diego in 2005, and he is a candidate for the Ph.D. degree, anticipating completion in 2008. At the University of San Diego, he has tackled problems in computer vision, including Object Recognition, Invariant Region Detection, Visual Tracking, and Head Pose Estimation. He has designed and implemented numerous real-time systems for human-computer interaction, intelligent environments, and driver assistance. Currently, he a Software Engineer at Google, Inc. in Mountain View, CA.

Mohan M. Trivedi received the B.E. degree (with honors) from the Birla Institute for Technology and Science in India, in 1974 and the Ph.D. degree from Utah State University, Logan, in 1979. He is a Professor of electrical and computer engineering and the founding Director of the Computer Vision and Robotics Research Laboratory, University of California at San Diego (UCSD), La Jolla, CA. He has a broad range of research interests in the intelligent systems, computer vision, intelligent ("smart") environments, intelligent vehicles and transportation systems, and human-machine interfaces areas. He has published nearly 300 technical articles and has edited over a dozen volumes, including books, special issues, video presentations, and conference proceedings.

Dr. Trivedi serves on the executive committee of the California Institute for Telecommunication and Information Technologies [Cal-IT2] as the Leader of the Intelligent Transportation and Telematics Layer at UCSD. He serves regularly as a Consultant to industry and government agencies in the USA and abroad. He was the Editor-in-Chief of the Machine Vision and Applications Journal. He served as the Chairman of the Robotics Technical Committee of the Computer Society of the IEEE and on the ADCOM of the IEEE SMC Society. He received the Distinguished Alumnus Award from the Utah State University, Pioneer Award (Technical Activities), and Meritorious Service Award from the IEEE Computer Society.

TABLE V

STANDARD DATASETS FOR HEAD POSE ESTIMATION

Pointing '04 – The Pointing '04 corpus [33] was included as part of the Pointing 2004 Workshop on Visual Observation of Deictic Gestures to allow for the uniform evaluation of head pose estimation systems. It was also used as one of two datasets to evaluate head pose estimators in the 2006 International Workshop on Classification of Events Activities and Relationships (CLEAR'06) [118]. The Pointing '04 consists of 15 sets of near-field images, with each set containing 2 series of 93 images of the same person at 93 discrete poses. The discrete poses span both pitch and yaw, ranging from -90° to 90° in both DOF. The subjects range in age from 20 to 40 years old, five possessing facial hair and seven wearing glasses. Each subject was photographed against a uniform background, and the heads were manually cropped. Head pose ground truth was obtained by directional suggestion. This led to poor uniformity between subjects, and as a result, the Pointing '04 dataset contains substantial errors. This has been noted in evaluations [125]. The data is publicly available at `http://www-prima.inrialpes.fr/Pointing04`.

CHIL-CLEAR06 – The CHIL-CLEAR06 dataset was collected for the CLEAR'06 workshop and contains multi-view video recordings of seminars given by students and lecturers [118]. The recordings consist of 12 training sequences and 14 test sequences recorded in a seminar room equipped with four synchronized cameras in each of the corners of the room. The length of the sequences range from 18 to 68 minutes each, and provide a far-field, low-resolution image of the head in natural indoor lighting. Manual annotations are made in every tenth video frame, providing a bounding box around the location of the presenter's head, along with a coarse pose orientation label from one of 8 discrete yaws at 45° intervals. More information about the dataset can be found at `http://www.clear-evaluation.org`.

CHIL-CLEAR07 – The CHIL-CLEAR07 dataset was created for the CLEAR'07 workshop and similar to CHIL-CLEAR06 contains multi-view video recordings of a lecturer in a seminar room [127]. In this iteration, however, head orientation was provided by a magnetic sensor, providing fine head pose estimates relative to a global coordinate system. The dataset consists of 15 videos with four-synchronized cameras captured at 15fps. In addition to the pose information from the magnetic sensor, which was capture for every frame, a manual annotation of the bounding box around each presenter's head was provided for every 5th frame. More information about the dataset can be found at `http://www.clear-evaluation.org`.

IDIAP Head Pose – The IDIAP Head Pose database consists of eight meeting sequences recorded from a single camera, each approximately one minute in duration [1]. It was provided as another source of evaluation for head pose estimators at the CLEAR'07 workshop, and from this context is often referred to as the **AMI** dataset. It consists of eight meeting sequences recorded from a single camera, each approximately one minute in duration. In each sequences, two subjects whom are always visible were continuously annotated using a magnetic sensor to measure head location and orientation. The head pose information is provided as fine measures of pitch and yaw relative to the camera, ranging within -90° and 90° for both DOF. The dataset is available on request, details can be found at `http://www.idiap.ch/HeadPoseDatabase`.

CMU PIE – The CMU Pose, Illumination, and Expression (PIE) dataset contains 68 facial images of different people across 13 poses, under 43 different illumination conditions, and with 4 different expressions [112]. The pose ground truth was obtained with a 13 cameras array, each positioned to provide a specific relative pose angle. This consisted of 9 cameras at approximately 22.5° intervals across yaw, one camera above the center, one camera below the center, and one in each corner of the room. Information regarding obtaining the dataset can be found at `http://www.ri.cmu.edu/projects/project_418.html`.

Softopia HOIP – The Softopia HOIP dataset consists of two sets of near-field face images of 300 people, (150 men and 150 women) against a uniform background [113]. The first set consists of 168 discrete poses (24 yaws and 7 pitches at 15° intervals), spanning frontal and rear views of each head. The second set contains finer yaw increments, capturing 511 discrete poses (73 yaws at 5° intervals and 7 pitches at 15° intervals). With both datasets, the images were captured with high precision, using camera arrays and rotating platforms. The dataset is available to Japanese academic institutions at `http://www.softopia.or.jp/rd/facedb.html`.

CVRR-86 – The CVRR-86 dataset contains 3894 near-field images of 28 subjects at 86 discrete poses against a uniform background [136]. Subjects were asked to move their heads while pose information was recorded with a magnetic sensor. Pitch and yaw were quantized into 15° intervals ranging from -45° to 60° and -90° to 90°, respectively. Not every pose has samples from each person, and some poses have more than one sample from each person. The dataset is not available online at this time, but more information can be found at `http://cvrr.ucsd.edu`.

CVRR-363 – The CVRR-363 dataset contains near-field images of 10 people against a uniform background [86]. 363 discrete poses ranging from -30° to 20° in pitch and -80° to 80° in yaw at 5° intervals were recorded with an optical motion capture system. To ensure that a single image was captured for each person at every discrete pose, the subjects were provided with visual feedback on multiple projection screens, and software ensured that an image was captured whenever the subject was with 0.5° of the desired pose. The dataset is not available online at this time, but more information can be found at `http://cvrr.ucsd.edu`.

USF HumanID – The USF HumanID Face dataset contains near-field color face images of 10 people against a uniform background [9]. There are 181 images of head person, varying across yaw from -90° to 90° in 1° increments. The dataset is not publicly available online at this time.

BU Face Tracking – The BU face tracking dataset contains two sets of 30fps video sequences with head pose recorded by a magnetic sensor [14]. The first set consists of 9 video sequences for each of five subjects taken under uniform illumination, and the second set consists of 9 video sequences for each of three subjects with time varying illumination. Subjects were asked to perform free head motion, which includes translations and rotations. The dataset is available online at `http://www.cs.bu.edu/groups/ivc/HeadTracking`.

CVRR LISAP-14 – The CVRR LISAP-14 dataset contains 14 video sequences of an automobile driver while driving in daytime and nighttime lighting conditions (eight sequences during the day, four sequences at night with near-IR illumination) [85]. Each sequence is approximately 8-minutes in length, and includes head position and orientation as recorded by an optical motion capture system. The dataset is not available online at this time, but more information can be found at `http://cvrr.ucsd.edu`.

CAS-PEAL The CAS-PEAL dataset contains 99,594 images of 1040 people with varying pose, expression, accessory, and lighting [29]. For each person, 9 yaws were simultaneously captured with a camera array at 3 pitches with directional suggestion, providing a total of 27 discrete poses. A subset of the data is available on request – more information can be found at `http://www.jdl.ac.cn/peal`.

FacePix The FacePix dataset contains 181 images for each of 30 subjects spanning -90° to 90° in yaw at 1° intervals [67]. The images were captured using a camera on a rotating platform and following the capture, they were cropped and scaled manually to ensure that the eyes and face appear at the same vertical position in every view. The dataset is available online at `http://cubic.asu.edu/resources/index.php`.