

# REAL-TIME HEAD POSE ESTIMATION USING A WEBCAM: MONOCULAR ADAPTIVE VIEW-BASED APPEARANCE MODEL

Louis-Philippe Morency  
USC Institute for Creative Technologies  
Marina del Rey, CA 90292  
morency@ict.usc.edu

## Abstract

Accurately estimating the person’s head position and orientation is an important task for a wide range of applications such as driver awareness and human-robot interaction. Over the past two decades, many approaches have been suggested to solve this problem, each with its own advantages and disadvantages. In this paper, we present a probabilistic framework called Monocular Adaptive View-based Appearance Model (MAVAM) which integrates the advantages from two of these approaches: (1) the relative precision and user-independence of differential registration, and (2) the robustness and bounded drift of keyframe tracking. In our experiments, we show how the MAVAM model can be used to estimate head position and orientation in real-time using a simple monocular camera. Our experiments on two previously published datasets show that the MAVAM framework can accurately track for a long period of time (>2 minutes) with an average accuracy of  $3.9^\circ$  and 1.2in with an inertial sensor and a 3D magnetic sensor.

## 1 Introduction

Real-time, robust head pose estimation algorithms have the potential to greatly advance the fields of human-computer and human-robot interaction. Possible applications include novel computer input devices (Fu and Huang, 2007), head gesture recognition, driver fatigue recognition systems (Baker et al., 2004), attention awareness for intelligent tutoring systems, and social interac-

tion analysis. Pose estimation may also benefit secondary face analysis, such as facial expression recognition and eye gaze estimation, by allowing the 3D face to be warped to a canonical frontal view prior to further processing.

Two main paradigms exist for automatically estimating head pose. *Dynamic* approaches, also called differential or motion-based approaches, track the position and orientation of the head through video sequences using pair-wise registration (i.e., transformation between two frames). Their strength is user-independence and higher precision for relative pose in short time scales, but they are typically susceptible to long time scale accuracy drift due to accumulated uncertainty over time. They also usually require the initial position and pose of the head to be set either manually or using a supplemental automatic pose detector. *keyframe-based* approaches, also called template-based approaches, use information previously acquired about the user (automatically or manually) to estimate the head position and orientation. These approaches are more accurate and suffer only bounded drift over time, but they lack the relative precision of dynamic approaches.

In this paper we present a Monocular Adaptive View-based Appearance Model (MAVAM) which integrates these two estimation paradigms described above in one probabilistic framework. The proposed approach has the high precision of a motion-based tracker and does not drift over time. MAVAM was specifically designed to estimate 6 degrees-of-freedom (DOF) of head pose in real-time from a single monocular camera with known internal calibration parameters (i.e., focal length and image center).

The following section describes previous work in head

# Report Documentation Page

*Form Approved  
OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE <b>DEC 2008</b>	2. REPORT TYPE <b>N/A</b>	3. DATES COVERED <b>-</b>	
4. TITLE AND SUBTITLE <b>Real-Time Head Pose Estimation Using A Webcam: Monocular Adaptive View-Based Appearance Model</b>		5a. CONTRACT NUMBER	
		5b. GRANT NUMBER	
		5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)		5d. PROJECT NUMBER	
		5e. TASK NUMBER	
		5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>USC Institute for Creative Technologies Marina del Rey, CA 90292</b>		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)	
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release, distribution unlimited</b>			
13. SUPPLEMENTARY NOTES <b>See also ADM002187. Proceedings of the Army Science Conference (26th) Held in Orlando, Florida on 1-4 December 2008, The original document contains color images.</b>			
14. ABSTRACT			
15. SUBJECT TERMS			
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>	<b>UU</b>
			18. NUMBER OF PAGES <b>8</b>
			19a. NAME OF RESPONSIBLE PERSON

pose estimation and explains the difference between MAVAM and other integration frameworks. Section 3 describes formally our view-based appearance model (MAVAM) and how it is adapted automatically over time. Section 4 explains the details of the estimation algorithms used to apply MAVAM to head pose tracking. Section 5 describes our experimental methodology and show our comparative results.

## 2 Previous Work

Over the past two decades, many techniques have been developed for estimating head pose. Very accurate shape models are possible using the Active Appearance Model (AAM) methodology (Cootes et al., 2001), such as was applied to 3D head data in (Blaiz and Vetter, 1999). However, tracking 3D AAMs with monocular intensity images is currently a time-consuming process, and requires that the trained model be general enough to include the class of the user being tracked.

Early work in the *dynamic* paradigm assumed simple shape models (e.g., planar (Black and Yacoob, 1995), cylindrical (La Cascia et al., 2000), or ellipsoidal (Basu et al., 1996)). Tracking can also be performed with a 3D face texture mesh (Schodl et al., 1998) or 3D face feature mesh (Wiskott et al., 1997). Some recent work looked morphable models rather than rigid models (Brand, 2001; Bregler et al., 2000; Torresani and Hertzmann, 2004). Differential registration algorithms are known for user-independence and high precision for short time scale estimates of pose change, but they are typically susceptible to long time scale accuracy drift due to accumulated uncertainty over time.

Some earlier work in *keyframe-based* paradigm include nearest-neighbors prototype methods (Wu and Trivedi, 2005; Fu and Huang, 2006) and template-based approaches (Kjeldsen, 2001). Vacchetti *et al.* suggested a method to merge online and offline keyframes for stable 3D tracking (Vacchetti et al., 2003). These approaches are more accurate and suffer only bounded drift over time, but they lack the relative precision of dynamic approaches.

Morency *et al.* (Morency et al., 2003) presented the Adaptive View-based Appearance Model (AVAM) for head tracking from stereo images. MAVAM generalizes the AVAM approach by operating on intensity im-

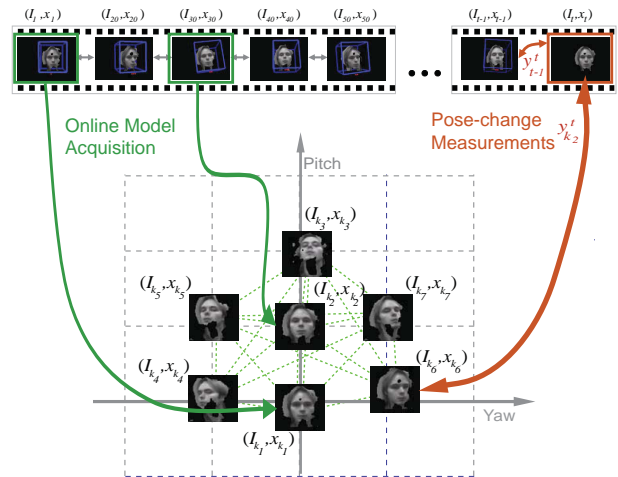


Figure 1: **Monocular Adaptive View-based Appearance Model (MAVAM)**. The pose of the current frame  $x_t$  is estimated using the pose-change measurements from two paradigms: differential tracking  $y_{t-1}^t$ , and keyframe tracking  $y_{k-2}^t$ . During the same pose update process (described in Section 3.3), the poses  $\{x_{k1}, x_{k2}, \dots\}$  from keyframes acquired online will be automatically adapted.

ages from a single monocular camera. This generalization faced two difficult challenges:

- Segmenting the face and selecting base frame set without any depth information by using a multiple face hypotheses approach (described in Section 3.1).
- Computing accurate pose-change estimation between two frames with only intensity images using iterative Normal Flow Constraint (described in Section 4.1);

MAVAM also includes some new functionality such as the keyframe management and a 4D pose tessellation space for the keyframe acquisition (see Section 3.4 for details). The following two sections formally describe this generalization.

### 3 Monocular Adaptive View-based Appearance Model

The two main components of the Monocular Adaptive View-based Appearance Model (MAVAM) are the view-based appearance model  $\mathcal{M}$  which is acquired and adapted over time, and the series of change-pose measurements  $\mathcal{Y}$  estimated every time a new frame is grabbed. Figure 1 shows an overview our MAVAM framework. Algorithm 1 presents a high-level overview of the main steps for head pose estimation using MAVAM.

A conventional view-based appearance model (Cootes et al., 2002) consists of different views of the same object of interest (e.g., images representing the head at different orientations). MAVAM extends the concept of view-based appearance model by associating a pose and covariance with each view. Our view-based model  $\mathcal{M}$  is formally defined as

$$\mathcal{M} = \{\{I_i, x_i\}, \Lambda_{\mathcal{X}}\}$$

where each view  $i$  is represented by  $I_i$  and  $x_i$  which are respectively the intensity image and its associated pose modeled with a Gaussian distribution, and  $\Lambda_{\mathcal{X}}$  is the covariance matrix over all random variables  $x_i$ . For each pose  $x_i$ , there exist a sub-matrix  $\Lambda_{x_i}$  in the diagonal of  $\Lambda_{\mathcal{X}}$  that represents the covariance of the pose  $x_i$ . The poses are 6 dimensional vector consisting of the translation and the three Euler angles  $[T^x \ T^y \ T^z \ \Omega^x \ \Omega^y \ \Omega^z]$ . The pose estimates in our view-based model will be adapted using the Kalman filter update with pose change measurements  $\mathcal{Y}$  as observations and the concatenated poses as the state vector. Section 3.3 describes this adaptation process in detail.

The views  $(I_i, x_i)$  represent the object of interest (i.e., the head) as it appears from different angles and depths. Different pose estimation paradigms will use different type of views:

- A differential tracker will use only two views: the current frame  $(I_t, x_t)$  and the previous frame  $(I_{t-1}, x_{t-1})$ .
- In a keyframe-based (or template-based) approach there will be  $1 + n$  views: the current frame  $(I_t, x_t)$  and the  $j = 1 \dots n$  keyframes  $\{I_{K_j}, x_{K_j}\}$ . Note that

---

**Algorithm 1** Tracking with a Monocular Adaptive View-based Appearance Model (MAVAM).

---

**for** each new frame  $(I_t)$  **do**

**Base Frame Set Selection:** Select the  $n_b$  most similar keyframes to the current current frame and add them to the base frame set. Always include the previous frame  $(I_{t-1}, x_{t-1})$  in the base frame set (see Section 3.1);

**Pose-change measurements:** For each base frame, compute the relative transformation  $y_s^t$ , and its covariance  $\Lambda_{y_s^t}$ , between the current frame and the base frame (see Sections 3.2 and 4 for details);

**Model adaptation and pose estimation:** Simultaneously update the pose of all keyframes and compute the current pose  $x_t$  by solving Equations 1 and 2 given the pose-change measurements  $\{y_s^t, \Lambda_{y_s^t}\}$  (see Section 3.3);

**Online keyframe acquisition and management:** Ensure a constant tessellation of the pose space in the view-based model by adding new frames  $(I_t, x_t)$  as keyframe if different from any other view in  $\mathcal{M}$ , and by removing redundant keyframes after the model adaptation (see Section 3.4).

**end for**

---

MAVAM acquires keyframes online and MAVAM adapts the poses of these keyframes during tracking so  $n$ ,  $\{x_{K_j}\}$  and  $\Lambda_{\mathcal{X}}$  change over time.

Since MAVAM integrates two estimation paradigms, its view-based model  $\mathcal{M}$  consists of  $2 + n$  views: the current frame  $(I_t, x_t)$ , the previous frame  $(I_{t-1}, x_{t-1})$ , and  $n$  keyframe views  $\{I_{K_j}, x_{K_j}\}$ , where  $j = 1 \dots n$ . The keyframes are selected online to best represent the head under different orientation and position. Section 3.4 will describe the details of this tessellation.

#### 3.1 Base Frame Set Selection

The goal of the base frame set into selection process is to find a subset of views (*base frames*) in the current view-based appearance model  $\mathcal{M}$  that are similar in appearance (and implicitly in pose) to the current frame  $I_t$ . This step reduces the computation time since pose-change measurements will be computed only on this subset.

To perform good base frame set selection (and pose-change measurements) we need to segment the face in the current frame. In the original AVAM algorithm (Morency et al., 2003), face segmentation was simplified by using the depth images from the stereo camera; with only an approximate estimate of the 2D position of the face and a simple 3D model of the head (i.e., a 3D box), AVAM was able to segment the face. Since MAVAM uses only a monocular camera model, its base frame set selection algorithm is necessarily more sophisticated. Algorithm 2 summarizes the base frame set selection process.

---

**Algorithm 2 Base Frame Set Selection** Given the current frame  $I_t$  and view-based model  $\mathcal{M}$ , returns a set of selected base frames  $\{I_s, x_s\}$ .

---

**Create face hypotheses for current frame** Based on the previous frame pose  $x_{t-1}$  and its associated covariance  $\Lambda_{x_{t-1}}$ , create a set of face hypotheses for the current frame (see Section 3.1 for details). Each face hypothesis is composed of a 2D coordinate and a scale factor representing the face center and its approximate depth.

**for** each keyframe  $(I_{K_j}, x_{K_j})$  **do**

**Compute face segmentation in keyframe** Position the ellipsoid head model (see Section 4.1) at pose  $x_{K_j}$ , back-project in image plane  $I_{K_j}$  and compute valid face pixels

**for** each current frames face hypothesis **do**

**Align current frame** Based on the face hypothesis, scale and translate the current image to be aligned with center of the keyframe face segmentation.

**Compute distance** Compute the L2-norm distance between keyframe and the aligned current frame for all valid pixel from the keyframe face segmentation.

**end for**

**Select face hypothesis** The face hypothesis with the smallest distance is selected to represent this keyframe.

**end for**

**Base frame set selection** Based on their correlation scores, add the  $n_b$  best keyframes in the base frame set. Note that the previous frame  $(I_{t-1}, x_{t-1})$  is always added to the base frame set.

---

The ellipsoid head model used to create the face mask for each keyframe is a half ellipsoid with the dimensions of an average head (see Section 4.1 for more details). The ellipsoid is rotated and translated based on the keyframe pose  $x_{K_j}$  and then projected in the image plane using the camera’s internal calibration parameters (focal length and image center).

The face hypotheses set represents different positions and scales of where the face could be in the current frame. The first hypothesis is created by projecting pose  $x_{t-1}$  from the previous frame in the image plane of the current frame. Face hypotheses are created around this first hypothesis based on the trace of the previous pose covariance  $tr(\Lambda_{x_{t-1}})$ . If  $tr(\Lambda_{x_{t-1}})$  is larger than a preset threshold, face hypotheses are created around the first hypothesis with increments of one pixel along both image plane axes and of 0.2 meters along the Z axis. Thresholds were set based on preliminary experiments and the same values used for all experiments. For each face hypothesis and each keyframe, a L2-norm distance is computed and the  $n_b$  best keyframes are then selected to be added in the base frame set. The previous frame  $(I_{t-1}, x_{t-1})$  is always added to the base frame set.

### 3.2 Pose-Change Measurements

Pose-change measurements are relative pose differences between the current frame and one of the other views in our model  $\mathcal{M}$ . We presume that each pose-change measurement is probabilistically drawn from a Gaussian distribution  $\mathcal{N}(y_s^t | x_t - x_s, \Lambda_{y_s^t})$ . By definition pose increments have to be additive, thus pose-changes are assumed to be Gaussian. Formally, the set of pose-change measurements  $\mathcal{Y}$  is defined as:

$$\mathcal{Y} = \{y_s^t, \Lambda_{y_s^t}\}$$

Different pose estimation paradigms will return different pose-change measurements:

- The differential tracker compute the relative pose between the current frame and the previous frame, and returns the pose change-measurements  $y_{t-1}^t$  with covariance  $\Lambda_{t-1}^t$ . Section 4.1 describes the view registration algorithm.

- The keyframe tracker uses the same view registration algorithm described in Section 4.1 to compute the pose-change measurements  $\{y_{K_s}^t, \Lambda_{y_{K_s}^t}\}$  between the current frame and the selected keyframes frames.

MAVAM integrates two estimation paradigms. Section 4 describes how the pose-change measurements are computed for head pose estimation.

### 3.3 Model Adaptation and Pose Estimation

To estimate the pose  $x_t$  of the new frame based on the pose-change measurements, we use the Kalman filter formulation described in (Morency et al., 2003). The state vector  $\mathcal{X}$  is the concatenation of the view poses  $\{x_t, x_{t-1}, x_{K_0}, x_{K_1}, x_{K_2}, \dots\}$  as described in Section 3 and the observation vector  $\mathcal{Y}$  is the concatenation of the pose measurement  $\{y_{t-1}^t, y_{K_0}^t, y_{K_1}^t, y_{K_2}^t, \dots\}$  as described in the previous section. The covariance between the components of  $\mathcal{X}$  is denoted by  $\Lambda_{\mathcal{X}}$ .

The Kalman filter update computes a prior for  $p(\mathcal{X}_t | \mathcal{Y}_{1..t-1})$  by propagating  $p(\mathcal{X}_{t-1} | \mathcal{Y}_{1..t-1})$  one step forward using a dynamic model. Each pose-change measurement  $y_s^t \in \mathcal{Y}$  between the current frame and a base frame of  $\mathcal{X}$  is modeled as having come from:

$$\begin{aligned} y_s^t &= C_s^t \mathcal{X} + \omega, \\ C_s^t &= \begin{bmatrix} I & 0 & \dots & -I & \dots & 0 \end{bmatrix}, \end{aligned}$$

where  $\omega$  is Gaussian and  $C_s^t$  is equal to  $I$  at the view  $t$ , equal to  $-I$  for the view  $s$  and is zero everywhere else. Each pose-change measurement  $(y_s^t, \Lambda_{y_s^t})$  is used to update all poses using the Kalman Filter state update:

$$[\Lambda_{\mathcal{X}_t}]^{-1} = [\Lambda_{\mathcal{X}_{t-1}}]^{-1} + C_s^{t\top} \Lambda_{y_s^t}^{-1} C_s^t \quad (1)$$

$$\mathcal{X}_t = \Lambda_{\mathcal{X}_t} \left( [\Lambda_{\mathcal{X}_{t-1}}]^{-1} \mathcal{X}_{t-1} + C_s^{t\top} \Lambda_{y_s^t}^{-1} y_s^t \right) \quad (2)$$

After individually incorporating the pose-changes  $(y_s^t, \Lambda_{y_s^t})$  using this update,  $\mathcal{X}_t$  is the mean of the posterior distribution  $p(\mathcal{M} | \mathcal{Y})$ .

### 3.4 Online Keyframe Acquisition and Management

An important advantage of MAVAM is the fact that keyframes are acquired online during tracking. MAVAM

generalized the previous AVAM (Morency et al., 2003) by (1) extending the tessellation space from 3D to 4D by including the depth of the object as the fourth dimension and (2) adding an extra step of keyframe management to ensure a constant tessellation of the pose space.

After estimating the current frame pose  $x_t$ , MAVAM must decide whether the frame should be inserted into the view-based model as a keyframe or not. The goal of the keyframes is to represent all different views of the head while keeping the number of keyframes low. In MAVAM, we use 4 dimensions to model the wide range of appearance. The first three dimensions are the three rotational axis (i.e., yaw, pitch and roll) and the last dimension is the depth of the head. This fourth dimension was added to the view-based model since the image resolution of the face changes when the user moves forward or backward and maintaining keyframes at different depths improves the base frame set selection.

In our experiments, the pose space is tessellated in bins of equal size: 10 degrees for the rotational axis and 100 millimeters for the depth dimension. These bin sizes were set to the pose differences that our pose-change measurement algorithm (described in Section 4.1) can accurately estimate.

The current frame  $(I_t, x_t)$  is added as a keyframe if either (1) no keyframe exists already around the pose  $x_t$  and its variance is smaller than a threshold, or (2) the keyframe closest to the current frame pose has a larger variance than the current frame. The variance of  $x_i$  is defined as the trace of its associated covariance matrix  $\Lambda_{x_i}$ .

The keyframe management step ensures that the original pose tessellation stays constant and no more than one keyframe represents the same space bin. During the keyframe adaptation step described in Section 3.3, keyframe poses are updated and some keyframes may have shifted from their original poses. The keyframe management goes through each tessellation bin from our view-based model and check if more than one keyframe pose is the region of that bin. If this is the case, then the keyframe with the lowest variance is kept while all the other keyframes are removed from the model. This process improves the performance of our MAVAM framework by compacting the view-based model.

## 4 Monocular Head Pose Estimation

In this subsection we describe in detail how the pose-change measurements  $y_s^t$  are computed for the different paradigms. For the differential and keyframe tracking,  $y_{t-1}^t$  and  $y_{K_j}^t$  are computed using Iterative Normal Flow Constraint described in the next section.

### 4.1 Monocular Iterative Normal Flow Constraint

Our goal is to estimate the 6-DOF transformation between a frame with known pose  $(I_s, x_s)$  and a new frame with unknown pose  $I_t$ . Our approach is to use a simple 3D model of the head (half of an ellipsoid) and an iterative version of the Normal Flow Constraint (NFC) (Vedula et al., 1999). Since pose is known for the base frame  $(I_s, x_s)$ , we can position the ellipsoid based on its pose  $x_s$  and use it to solve the NFC linear system. The Algorithm 3 shows the details of our iterative NFC.

## 5 Experiments

The goal is to evaluate the accuracy and robustness of the MAVAM tracking framework on previously published datasets. The following section describes these datasets while Section 5.2 presents the details of the models compared in our experiments. Our results are shown in Sections 5.3 and 5.4. Our C++ implementation of MAVAM runs at 12Hz on one core of an Intel X535 Quad-core processor. The system was automatically initialized using the static pose estimator described in the previous section.

### 5.1 Datasets

We evaluated the performance of our approach on two different datasets: the BU dataset from La Cascia *et al.* (La Cascia et al., 2000) and the MIT dataset from Morency *et al.* (Morency et al., 2003).

**BU dataset** consists of 45 sequences (nine sequences for each of five subjects) taken under uniform illumination where the subjects perform free head motion including translations and both in-plane and out-of-plane rotations. All the sequences are 200 frames long (approximately seven seconds) and contain free head motion of several

---

**Algorithm 3 Iterative Normal Flow Constraint** Given the current frame  $I_t$ , a base frame  $(I_s, x_s)$  and the internal camera calibration for both images, returns the pose-change measurement  $y_s^t$  between both frames and its associated covariance  $\Lambda_{y_s^t}$ .

---

**Compute initial transformation** Set initial value for  $y_s^t$  as the 2D translation between the face hypotheses for the current frame and the base frame (see Section 3.1

**Texture the ellipsoid model** Position the ellipsoid head model at  $x_s + y_s^t$ . Map the texture from  $I_s$  on the ellipsoid model by using the calibration information

**repeat**

**Project ellipsoid model** Back-project the textured ellipsoid in the current frame using the calibration information.

**Normal Flow Constraint** Create a linear system by applying the normal flow constraint (Vedula et al., 1999) to each valid pixel in the current frame.

**Solve linear system** Estimate  $\Delta_{y_s^t}$  by solving the NFC linear system using linear least square. Update the pose-change measurement  $y_s^{t(new)} = y_s^{t(old)} + \Delta_{y_s^t}$  and estimate the covariance matrix  $\Lambda_{y_s^t}$  (Lawson and Hanson, 1974).

**Warp ellipsoid model** Apply the transformation  $\Delta_{y_s^t}$  to the ellipsoid head model

**until** Maximum number of iterations reached or convergence:  $trace(\Lambda_{y_s^t}) < T_\Lambda$

---

subjects. Ground truth for these sequences was simultaneously collected via a ‘‘Flock of Birds’’ 3D magnetic tracker (??, flock). The video signal was digitized at 30 frames per second at a resolution of 320x240. Since the focal length of the camera is unknown, we approximated it to 500 (in pixel) by using the size of the faces and knowing that they should be sitting approximately one meter from the camera. This approximate focal length add challenges to this dataset. **MIT dataset** contains 4 video sequences with ground truth poses obtained from an *Inertia Cube*<sup>2</sup> sensor. The sequences were recorded at 6 Hz and the average length is 801 frames (~133sec). During recording, subjects underwent rotations of about 125 degrees and translations of about 90cm, including translation along the Z axis. The sequences were originally recorded using a stereo camera from Videre Design (De-

Technique	Tx	Ty	Tz
MAVAM	1.00in	0.88in	1.82in
Technique	Pitch	Yaw	Roll
MAVAM	3.73°	5.44°	2.79°

Table 1: Average accuracies on BU dataset (La Cascia et al., 2000). MAVAM successfully tracked all 45 sequences while La Cascia *et al.* (La Cascia et al., 2000) reported an average percentage of tracked frame of only  $\sim 75\%$ .

sign, 2000). For our experiments, we used only the left images. The exact focal length was known. By sensing gravity and earth magnetic field, *Inertia Cube*<sup>2</sup> estimates for the axis X and Z axis (where Z points outside the camera and Y points up) are mostly driftless but the Y axis can suffer from drift. InterSense reports a absolute pose accuracy of 3°RMS when the sensor is moving. This dataset is particularly challenging since the recorded frame rate was low and so the pose differences between frames will be larger.

## 5.2 Models

We compared two models for head pose estimation: our approach MAVAM as described in this paper, and the original stereo-based AVAM (Morency et al., 2003).

**MAVAM** The Monocular Adaptive View-based Appearance Model (MAVAM) is the complete model as described in Section 3. This model integrates two pose estimation paradigms: differential tracking and keyframe tracking. It is applied on monocular intensity images.

**3D AVAM** The stereo-based AVAM is the original model suggested by Morency *et al.* (Morency et al., 2003). The results for this model are taken directly from their research paper. Since this model uses intensity images as well as depth images, we should expect better accuracy for this 3D AVAM.

## 5.3 Results with BU dataset

The BU dataset presented in (La Cascia et al., 2000) contains 45 video sequences from 5 different people. The results published by La Cascia *et al.* are based on three error criteria: the average % of frames tracked, the po-

Technique	Pitch	Yaw	Roll
MAVAM	5.3° ± 15.3°	4.9° ± 9.6°	3.6° ± 6.3°
3D AVAM	2.4°	3.5°	2.6°

Table 2: Average rotational accuracies on MIT dataset (Morency et al., 2003). MAVAM performs almost as well as the 3D AVAM which was using stereo calibrated images while our MAVAM works with monocular intensity images.

sition error and the orientation error. The position and orientation errors includes only the *tracked* frames and ignores all frames with very large error. In our results, the MAVAM successfully tracked all 45 video sequences without losing track at any point. The Table 1 shows the accuracy of our MAVAM pose estimator. The average rotational accuracy is 3.9° while the average position error is 1.2inches (3.1cm). These results show that MAVAM is accurate and robust even when the focal length can only be approximated.

## 5.4 Results with MIT dataset

The MIT dataset presented in (Morency et al., 2003) contains four long video sequences ( $\sim 2$ mins) with a large range of rotation and translation. Since the ground truth head positions were not available for this dataset, we present results for pose angle estimates only. Table 2 shows the averaged angular error the different models. The results for 3D AVAM were taken for the original publication (Morency et al., 2003). We can see that MAVAM performs almost as well as the 3D AVAM which was using stereo calibrated images while our MAVAM works with monocular intensity images.

## 6 Conclusion

In this paper, we presented a probabilistic framework called Monocular Adaptive View-based Appearance Model (MAVAM) which integrates the advantages from three of these approaches: (1) the relative precision and user-independence of differential registration, and (2) the robustness and bounded drift of keyframe tracking. On two challenging 3-D head pose datasets, we demonstrated



that MAVAM can reliably and accurately estimate head pose and position using a simple monocular camera.

## Acknowledgements

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the U.S. Naval Research Laboratory under grant # NRL 55-05-03. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## References

- Baker, S., I. Matthews, J. Xiao, R. Gross, T. Kanade, and T. Ishikawa (2004). Real-time non-rigid driver head tracking for driver mental state estimation. In *Proc. 11th World Congress Intelligent Transportation Systems*.
- Basu, S., I. Essa, and A. Pentland (1996). Motion regularization for model-based head tracking. In *Proceedings. International Conference on Pattern Recognition*.
- Black, M. and Y. Yacoob (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV*, pp. 374–381.
- Blanz, V. and T. Vetter (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH99*, pp. 187–194.
- Brand, M. (2001). Morphable 3D models from video. In *CVPR*.
- Bregler, C., A. Hertzmann, and H. Biermann (2000). Recovering non-rigid 3D shape from image streams. In *CVPR*.
- Cootes, T., G. Edwards, and C. Taylor (2001, June). Active appearance models. *PAMI* 23(6), 681–684.
- Cootes, T. F., G. V. Wheeler, K. N. Walker, and C. J. Taylor (2002, August). View-based active appearance models. *Image and Vision Computing Volume 20*, 657–664.
- Design, V. (2000). *MEGA-D Megapixel Digital Stereo Head*. <http://www.ai.sri.com/konolige/svs/>.
- flock. *The Flock of Birds*. P.O. Box 527, Burlington, Vt. 05402.
- Fu, Y. and T. Huang (2006). Graph embedded analysis for head pose estimation. In *Proc. IEEE Intl. Conf. Automatic Face and Gesture Recognition*, pp. 3–8.
- Fu, Y. and T. S. Huang (2007). hmouse: Head tracking driven virtual computer mouse. In *IEEE Workshop Applications of Computer Vision*, pp. 30–35.
- Kjeldsen, R. (2001). Head gestures for computer control. In *Proc. Second International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pp. 62–67.
- La Cascia, M., S. Sclaroff, and V. Athitsos (2000, April). Fast, reliable head tracking under varying illumination: An approach based on registration of textured-mapped 3D models. *PAMI* 22(4), 322–336.
- Lawson, C. and R. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall.
- Morency, L.-P., A. Rahimi, and T. Darrell (2003). Adaptive view-based appearance model. In *CVPR*, Volume 1, pp. 803–810.
- Schodl, A., A. Haro, and I. Essa (1998). Head tracking using a textured polygonal model. In *PUI98*.
- Torresani, L. and A. Hertzmann (2004). Automatic non-rigid 3D modeling from video. In *ECCV*.
- Vacchetti, L., V. Lepetit, and P. Fua (2003). Fusing online and offline information for stable 3d tracking in real-time. In *CVPR*.
- Vedula, S., S. Baker, P. Rander, R. Collins, and T. Kanade (1999). Three-dimensional scene flow. In *ICCV*, pp. 722–729.
- Wiskott, L., J. Fellous, N. Kruger, and C. von der Malsburg (1997, July). Face recognition by elastic bunch graph matching. *PAMI* 19(7), 775–779.
- Wu, J. and M. M. Trivedi (2005). An integrated two-stage framework for robust head pose estimation. In *International Workshop on Analysis and Modeling of Faces and Gestures*.