Inferring Body Pose using Speech Content

Sy Bor Wang
MIT CSAIL
70 Vassar St
Cambridge, MA, USA
sybor@csail.mit.edu

David Demirdjian
MIT CSAIL
70 Vassar St
Cambridge, MA, USA
demirdji@csail.mit.edu

ABSTRACT

Untethered multimodal interfaces are more attractive than tethered ones because they are more natural and expressive for interaction. Such interfaces usually require robust vision-based body pose estimation and gesture recognition. In interfaces where a user is interacting with a computer using speech and arm gestures, the user's spoken keywords can be recognized in conjuction with a hypothesis of body poses. This co-occurence can reduce the number of body pose hypothesis for the vision based tracker. In this paper we show that incorporating speech-based body pose constraints can increase the robustness and accuracy of vision-based tracking systems.

Next, we describe an approach for gesture recognition. We show how Linear Discriminant Analysis (LDA), can be employed to estimate 'good features' that can be used in a standard HMM-based gesture recognition system. We show that, by applying our LDA scheme, recognition errors can be significantly reduced over a standard HMM-based technique.

We applied both techniques in a *Virtual Home Desktop* scenario. Experiments where the users controlled a desktop system using gestures and speech were conducted and the results show that the speech recognised in conjunction with body poses has increased the accuracy of the vision-based tracking system.

Categories and Subject Descriptors

H.1.2 [**User-Machine System**]: Human Information processing; I.5.4 [**Computing Methodologies**]: Pattern Recognition Applications—*computer vision*

General Terms

Algorithms, Experimentation

Keywords

Audio-Visual Tracking, Untethered Body Pose Tracking, Arm Gesture Recogntion

ICMI'05, October 4-6, 2005, Trento, Italy

1. INTRODUCTION

Multimodal interfaces have gained a lot of attention in research ever since Bolt's Put-That-There system.[1]. Understanding untethered arm gestures in multimodal interfaces is very useful as it allows richer expression and more natural interaction from users. Robust 3D body pose estimation and gesture recognition techniques, however, are required.

Many 3D gesture systems using a vision-based tracker encounter tracking errors, leading to multiple interpretations in gestures. In many existing multimodal interfaces, speech has been used to improve gesture recognition. Previous work in a pen-input system[12] and a virtual reality system[9] have used speech to disambiguate gestures. User studies in pen-input systems [14] and in 3D gestures systems[5, 2], have shown that speech and gestures are correlated in time. These systems have shown consistently that the start point of a gesture is very closely correlated to the start time of the deictic word that expresses the same intent. Weather narration[17] has employed the use of this co-occurrence and successfully improved gesture recognition. Since this co-occurrence has been effective in improving gesture recognition, we question if the same insight could be used to improve tracking in a vision based-tracker. Some vision-based tracker generates multiple hypothesis of body poses. In a human-computer interface which involves arm gestures and speech,some of these body poses occur in conjunction with the user's spoken words in a consistent fashion. This presence of the spoken word can help reduce the number of hypothesised poses generated by the tracker.

In this paper, our primary goal is to show that the robustness and accuracy of a vision-based tracking system can be increased by considering body pose constraints induced by the speech content. We show that our technique estimates more accurate deictic references when speech is available.

Besides improving the accuracy of the vision-based tracker, the robustness of multimodal interfaces can improved by gesture recognition as well. As a secondary goal of this paper, we present an approach for gesture recogntion. Using the concept of phonemes in speech recognition, we show how Linear Discriminant Analysis (LDA) can be employed to estimate some 'good features' that can be used in a standard HMM-based gesture recognition system. We applied our LDA approach, and showed that recognition errors can be decreased significantly over a standard HMM-based technique.

Finally, we applied the integration of our multimodal approach for tracking and gesture recognition in a *Virtual Home Desktop* application. This application allows a user to

manipulate windows and programs on a desktop projected on a wall.

2. RELATED WORK

Ever since Bolt's Put-That There[1] System, many other 3D multimodal interfaces have emerged. Koons et al [10] presented a system that tracked 3D hand-based pointing gestures, speech , and gaze, but did not use speech information to assist in tracking of the hand. Corradini[2] augmented Quickset[14], a multimodal voice/pen system that allows users to create and control maps, such that it could accept 3D hand movements as user input. He showed how the system supported the concept of mutual disambiguation. Sharma[17], conducted a co-occurrence analysis of a selected set of spoken keywords with different gestures to improve the performance of a HMM-based gesture recognizer. More recently, a human-robot interface[5, 18] in Germany used speech to resolve ambuiguity in deictic gestures. All these systems have used speech to improve gesture recognition or deictic resolution. In our work, we would extend this further by using speech to improve body pose estimation.

3. TRACKING WITH SPEECH CONTENT

This section briefly describes our real-time model-based tracking algorithm. Our algorithm can be considered as an audio-visual extension of our previous work on vision-based tracking [4].

3.1 Fitting Error

We consider the pose estimation problem as the fitting of a body model pose $\mathbf{\Pi}$ to a set of visual observations. In this work, $\mathbf{\Pi}$ consists of the relative orientation $\boldsymbol{\theta}_i$ $(i \geq 1)$ between consecutive limbs.

When visual observations come from a stereo or multiview camera, tridimensional reconstructions $\mathcal{P} = \{M_i\}$ of the points M_i in the scene can be estimated. In this case, a fitting error function $E(\mathbf{\Pi})$ defined as the distance between reconstructed points $\mathcal{P}$ and the 3D model at pose $\mathbf{\Pi}$ is suitable. Such a function can be defined as:

$$E^2(\mathbf{\Pi}) = \sum_{M_i \in \mathcal{P}} d^2(M_i, \mathcal{B}(\mathbf{\Pi})) \quad (1)$$

where $\mathcal{B}(\mathbf{\Pi})$ is the 3D reconstruction of the body model at pose $\mathbf{\Pi}$ and $d^2(M_i, \mathcal{B}(\mathbf{\Pi}))$ the Euclidean distance between the point M_i and the 3D model $\mathcal{B}(\mathbf{\Pi})$.

3.2 Motion Constraints

A direct approach commonly used for pose tracking [3, 8, 4] consists of performing a local minimization of the fitting error $E(\mathbf{\Pi})$ using the pose $\mathbf{\Pi}_{t-1}$ estimated at the previous frame as initialization.

However, estimating the pose $\mathbf{\Pi}$ by minimizing the fitting error $E(\mathbf{\Pi})$ only is usually not enough to provide a robust tracking. Indeed, the minimization of $E(\mathbf{\Pi})$ is likely to fall into local minima, causing the tracking to fail. This happens, for instance, when users perform fast motion (the minimization starts with an initialization far from the true minimum) or in case of partial occlusion of the body (corresponding to ambiguities in the pose).

The robustness of the tracking system can be increased by adding constraints of the body pose $\mathbf{\Pi}$. Indeed, additional constraints on the body pose $\mathbf{\Pi}$ (or equivalently on the body motion) can reduce the search space to the 'good' directions, therefore significantly improving the tracking robustness. For instance, when tracking a user pointing at a screen, multiple constraints on the body pose (e.g. user in a standing position, torso facing in the direction of the screen, one arm pointing toward the screen) seem natural to impose. In a similar manner, when the user is talking to a person or is referring to a visible object in the scene, he is more likely to face the person or object. Therefore tracking the user using a body model constrained to have the torso facing the physical referents is appropriate.

In this paper, we define the context c as the type of physical action, task or gesture a user is performing (e.g. pointing at the screen, showing an object, talking to someone, resting, agreeing/disagreeing). Let $\mathcal{C}_c$ be a set of motion constraints associated with a context c. In this work, constraints are expressed using fixed bounds for the angles between the axes of consecutive limbs or for the relative orientation of a limb with respect to the world coordinate system. These constraints can be written as a set of inequalities on the elements of $\mathbf{\Pi}$ such as:

$$|\Pi(i)| \leq angle_i^{(c)}$$

If the context c in which the user is is known, an estimation of the pose $\mathbf{\Pi}^{(c)}$ of the user is performed by minimizing eq.(1) subject to the constraints $\mathcal{C}_c$.

Next we explore the use of speech to provide some context information (when applicable) about the pose and motion of the user. The context extracted from the speech is then used to provide a set of pose and motion constraints to the tracking algorithm which concurrently evaluates the different hypotheses.

3.3 Using Speech to Provide Physical Constraints

Here we describe how speech content is used to generate motion models hypotheses. The co-occurrence of words and gestures is used to provide motion models.

Let W and probabilities $p(W)$ be respectively a set of words and the probabilities that they have been pronounced. Let $p(c|W)$ be the probability of observing the context c conditioned on the word W being spoken. The conditional probabilities $p(c|W)$ are empirically determined from a user study described in Section 4.2.

The probability of observing the context c given speech observation is:

$$p(c) = \sum_W p(c|W)p(W) \quad (2)$$

As part of our model, we introduce a non-context $c = 0$ corresponding to the absence of specific context. For words W which do not co-occur with any specific gestures, the conditional probability $p(c = 0|W)$ is close to 1.

3.4 Audio-Visual Tracking

The complete audio-visual tracking algorithm consists in concurrently estimating the poses corresponding to the most probable contexts. More precisely, the contexts c with highest probability $p(c)$ are evaluated from audio observations (speech recognizer) and used to evaluate $\mathbf{\Pi}^{(c)}$ by constrained optimization of eq.(1) subject to the constraints $\mathcal{C}_c$.

The final pose Π is estimated as a weighted sum of the

contextual poses $\mathbf{\Pi}^{(c)}$:

$$p(\mathbf{\Pi}) = \sum_c w(\mathbf{\Pi}^{(c)})\mathbf{\Pi}^{(c)} \quad (3)$$

where the weights $w(\mathbf{\Pi}^{(c)})$ are function of the fitting error eq. (1).

It is important to notice that our approach is robust to errors in speech. Indeed, if a word is mis-detected with a high probability and induced a wrong context c, the weight $w(\mathbf{\Pi}^{(c)})$ of the pose $\mathbf{\Pi}^{(c)}$ will be small (because of the inadequacy between the motion model and the visual data).

3.5 User Study

Five English speakers participated in a user study to observe their multimodal interaction with a simulated "Home Desktop System" application. This application basically allows users to control windows or programs on the computer screen without using the keyboard or mouse. The computer screen is projected on a wall and the users have to stand in front of it to control the system.

First, the five speakers were given an orientation about the various objects for manipulation on the computer screen. Then, they were briefed on a set of gestures they can use in controlling these objects. Next, they were given various tasks in controlling these objects. These tasks include resizing a particular window, re-positioning it and minimizing it etc.. These users were free to use either or both speech and hand gestures to perform their tasks. They were encouraged to speak naturally with the gestures, to work at their own pace, and to focus on completing their tasks. While they are performing these tasks, their gestures and speech were videotaped and recorded. The goal of this study is to find out if there is a common set of words associated with a set of body poses. The age of all five users range between 21 and 33 years old. Figure 1 shows a sample image of a user taking the study.

Our study found that certain pose context occur consistently when specific words were spoken. From the study, we determined the empirical probability of the pose context given the spoken word, $p(c|W)$. Table 1 shows a sample set of pose contexts associated with a spoken word and their probabilities. A total of 29 keywords were found to have consistent association with certain pose contexts. We will use these results for a user experiment (described in Section 5.1) to evaluate the performance of our audio-visual tracking algorithm.

Word/s, W	**Context,** c	$p(c\|W)$
Backward	Both hands stretched in front	0.75
	Other	0.25
Click	Pointing	0.88
	Other	0.12
Down	Pointing	0.73
	Right Hand Raised	0.18
	Other	0.09
Fill	Both hands stretched laterally	0.33
	Both hands stretched vertically	0.33
	Other	0.33
OK	Both hands stretched in front	0.57
	Pointing	0.14
	Other	0.29
Stop	Both hands stretched in front	0.31
	Pointing	0.69
Up	Waving	0.40
	Pointing	0.60

Table 1: Sample set of spoken keywords, their associated pose context and their posterior probabilities

4. LINEAR DISCRIMINANT ANALYSIS FOR GESTURE RECOGNITION

Using the body poses, $\mathbf{\Pi}$ from the audio-visual tracking algorithm described in Section 3, we estimated features which will be fed to a Hidden Markov Model(HMM)[15] for gesture recognition. Hidden Markov Models (HMMs), usually perform well when the number of gestures to recognize is small. However, their performance usually decreases tremendously as the number and similarity of the gestures grows. One of the reasons is that the features of different gestures are too close to one another in the feature space. To overcome this, we show how Linear Discriminant Analysis (LDA), a technique used in feature separation of phonemes in speech recognition [6], can improve the performance of an arm gesture recognition system.

We propose modelling body gestures using elementary units, called "gestemes"[7]. Gestemes are estimated in the following way.

An input feature vector, f_n is a vector of K consecutive poses.

$$f_n = (\mathbf{\Pi_n}, \mathbf{\Pi_{n+1}}, ..., , \mathbf{\Pi_{n+K-1}})^T$$

Such a feature vector incorporates information about the pose and motion dynamics. We assume that there are B gestemes that models all the features in our L classes of gestures. A Gaussian Mixture Model containing B centers was estimated using a set of training feature vectors. All the Gaussians are assumed to have full covariances. Parameters for the B Gaussians are determined from the set of training feature vectors $f_n, 1 \leq n \leq N$, using the EM algorithm to maximize a global likelihood function given below:

$$L = \sum_{n=1}^{N} log \sum_{i=1}^{B} \delta_i p(f_n|i)$$

where $p(f_n|i)$ is a single component of the mixture of Gaussians, and δ_i is the ith component's mixing proportion.

Once this likelihood is maximized, each Gaussian cluster is assigned a unique label. The probabilities of each training feature belonging to each cluster, $p(i|f_n)$ are evaluated to determine the most likely cluster they belong to.

$$L_{f_n} = \arg\max_i p(i|f_n)$$

where L_{f_n} is the best scoring Gaussian's label assigned to training feature f_n. After this classification of the feature vectors into gesteme clusters, a projection matrix, that separates these feature vectors in an optimal way, is estimated by running LDA on these labelled training features.

Suppose there are B labels, then the within class expected covariance is:

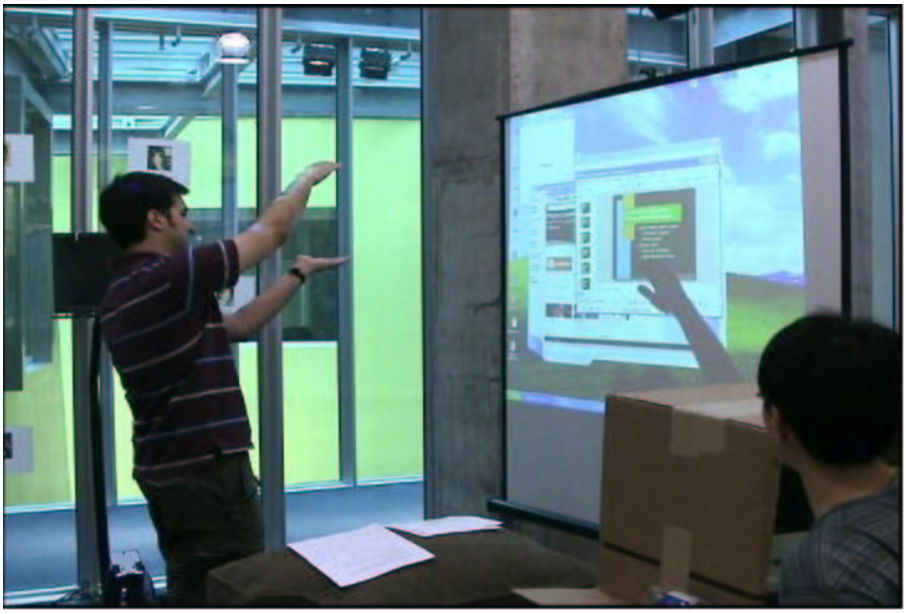

Figure 1: User gesturing in front of the projected desktop system

$$S_w = \sum_{i=1}^{B} p_i \cdot \Lambda_i$$

where p_i is the prior probability and Λ_i is the covariance of Gaussian i respectively.

The mean of the whole mixture of Gaussians is computed as

$$u_{overall} = \sum_{i=1}^{B} p_i \cdot u_i$$

where u_i is the mean of each Gaussian component.

The between class variance is then computed as

$$S_b = \sum (u_j - u_{overall}) \times (u_j - u_{overall})^T$$

Assuming a class-independent transform, the optimizing criterion, $J(\boldsymbol{w})$ is computed as

$$J(\boldsymbol{w}) = (S_w)^{-1} \times S_b$$

The projection matrix, $\mathbf{\Gamma}$, is found as the eigenvector matrix of $J(\boldsymbol{w})$ and the projected feature vector, $f_n^{'}$ becomes

$$f_n^{'} = \mathbf{\Gamma}^T f_n$$

These projected feature vectors, which are more distinct between classes, and less dispersed within each class, are then used for training the HMM to classify the features by the gesture classes.

5. EXPERIMENTS

5.1 Speech Constrained Tracking

We conducted a set of user experiments to evaluate the performance of our Audio-Visual tracking algorithm described in Section 3.4. This time, nine users participated in the experiment. In a similar setting as the user study described in Section 3.5, users were given an orientation about various icons or windows for manipulation on the computer screen. Then, they were briefed on a set of gestures to use for controlling these objects. Next they were given a few tasks to perform. During the experiments, the ground truth data corresponding to the location of the objects deictically referred and the gesture performed was manually transcribed.

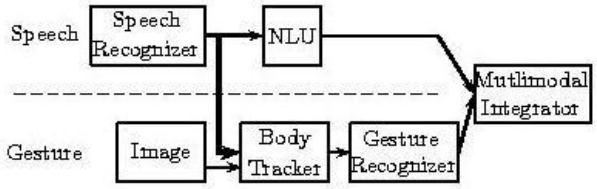

Figure 2: System Diagram

While the users were performing these tasks, the audio visual data was used to track the user and perform the gesture recognition. Color and stereo images of the user were captured using a standard stereo camera. The users wore a head-mounted microphone and their speech was sent to the GALAXY[16] system for speech recognition. The GALAXY system was trained with domain-specific grammars. For each spoken utterance, the GALAXY Speech Recognizer generates a time-stamped N-best list of words, and a probability, $p(W)$ is assigned to each hypothesized word. The probability of the context c is then computed using eq. (2). Finally, the algorithm described in Section 3 is used to estimate the user's pose.

In these experiments, the Word Error Rate of the Speech Recognizer was 37.02%. The interactions between the various sub-systems are shown in Figure 2.

We compared the performance of our system (depicted in

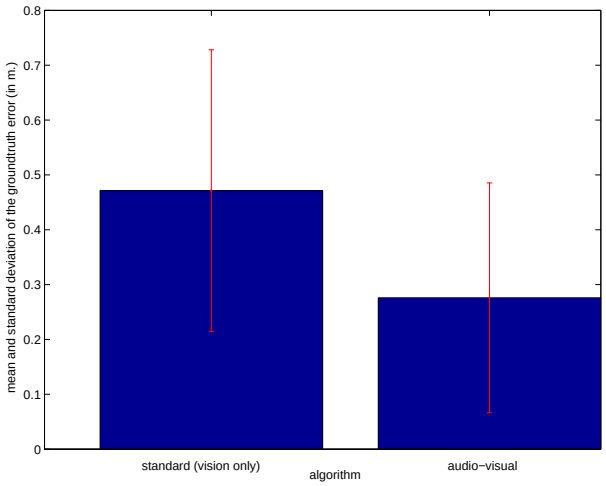

Figure 3: Average ground truth error for the pointing estimation. The average error for our audio-visual approach is about 0.27 $m.$ (corresponding to about 7 $deg.$) compared to about 0.47 $m.$ (corresponding to about 13 $deg.$) for the standard vision-based tracking algorithm.

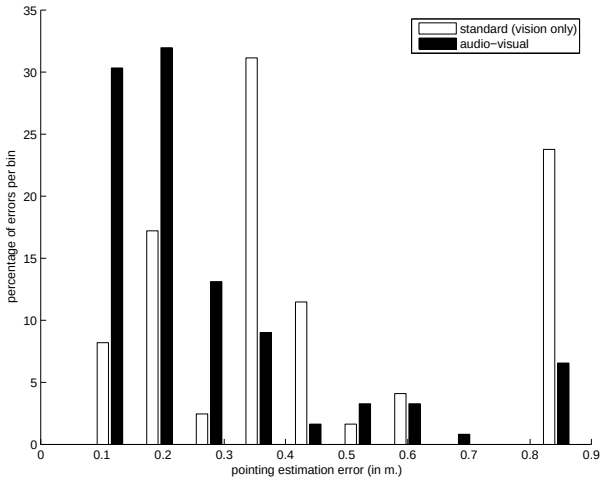

Figure 4: Histogram of the ground truth error for the pointing estimation.

Figure 2) to a standard (context-independent) vision-based tracker.

The location of the point of the screen pointed at by the user was computed as the intersection of the line formed by the (right) shoulder and hand of the estimated 3D model and the plan of the screen.

First, the tracking performance of the algorithms was evaluated. In absence of ground truth for the pose of the users, the pose provided by the tracking algorithms was visually inspected in the testing sequences. The standard tracker (vision only) had an error rate of about 20%. The audio-visual tracker had an error rate of about 12%.

Then, we evaluated the aptitude of the algorithms to provide correct tracking information when the users were performing deictic references. Figure 3 shows the average ground truth error (in $m.$) for the pointing estimation. Figure 4 shows a more detailed distribution of the error. The results show that our audio-visual approach for tracking has less outliers (large errors) and provides more accurate pointing locations than the standard tracker. The average error for our audio-visual approach is about 0.27 $m.$ (corresponding to about 7 $deg.$) compared to about 0.47 $m.$ (corresponding to about 13 $deg.$) for the standard vision-based tracking algorithm.

Figure 5 shows the variation of the fitting error $E^2(\mathbf{\Pi})$ during one of the sequences used in our experiments. The graph shows that around frames 1850, 1980 and 2020, the use of context from speech contributes a fitting error smaller than the vision-based only tracking system (which fails to track the user correctly). Figure 6 shows some tracking results corresponding to the frame 1860 of this sequence. Around this frame, the user pronounced the word *'move'*, which induced a *pointing context* (right arm pointing, torso facing the projection screen) that was used in the audio-visual tracking algorithm. The vision-based only tracker actually failed to track the user correctly at this frame and corresponds to the large fitting error for this frame in the Figure 5.

5.2 Gesture Recognition Results

We labeled eleven gestures for the HMM-based recognizer to classify. Gesture data was collected and gathered from thirteen users. These users were asked to perform the gestures in front of a stereo camera, and their body pose sequences of these gestures were collected. Half of these samples were used for training the HMM recognizer and for linear discriminant analysis, while the other half was used for testing. Examples of these gestures are shown in Figure 7.

We experimented with different techniques or different types of features, while varying the feature length (the feature length is the number of consecutive poses used to form each feature vector), and our results are plotted in Figure 8. In the first technique, we applied LDA to pose features and sent the projected features, $f_n^{'} = \mathbf{\Gamma}^T f_n$ to train the HMM (The plot from this technique is labelled LDA Features in the figure). In the second technique, we simply used pose features, f_n to train the HMM (labelled as pose features in the figure). In the third technique, we used velocity from consecutive poses as features for training the HMM (labelled as velocity features, ν_n). A velocity feature is given as

$$\nu_n = f_n - f_{n-1}$$

In the fourth technique, we simply used both pose and velocity features for training the HMM (labelled as velocity and pose features). A velocity and pose feature is given as

$$\psi_n = \begin{pmatrix} f_n \\ \nu_n \end{pmatrix}$$

In the fifth technique, we applied principal component analysis on the pose features (labelled as PCA). The components that contribute to 98% of the energy were extracted, and the projected pose features were used to train the HMM. In the last technique, we applied kernel principal component analysis (kPCA) on the pose features, and used the projected features to train the HMM. We used a Gaussian ker-

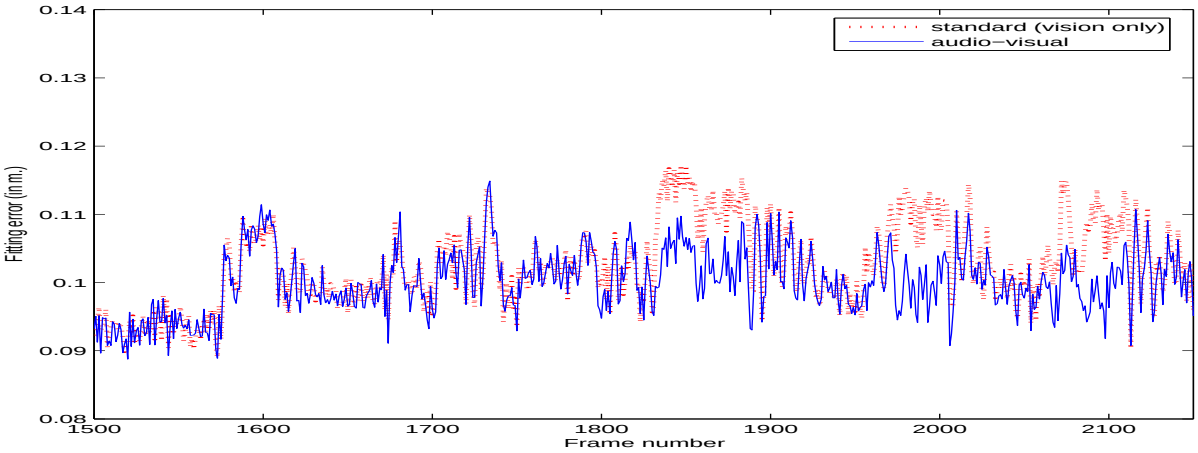

Figure 5: Average fitting error E between the estimation of the 3D articulated model and the 3D scene reconstruction *vs.* number of frames. Peaks in the data (around frames 1850, 1980 and 2020) correspond to tracking failures of the standard (vision only) tracking algorithm.

nel for the projection. The Gaussian kernel function $k(x, y)$ is given by:

$$k(x, y) = e^{-\frac{||x-y||^2}{2 \cdot \sigma^2}}$$

and we used a σ value of 10. We did not display results using other kernel functions as they performed significantly worse.

From Figure 8, pose features projected into the LDA subspace performed consistently better than the other 5 techniques. The error rate of the kPCA technique did not perform as well as we expected. We are currently investigating this matter.

We trained the HMM 50 times using the various techniques or feature types at a constant feature length of 4, and tabulated the statistics of the error rates. Our results are given in Table 2. As a comparison to pose features, the error rate was divided by a factor of 1.5 by applying LDA. The standard deviation of the error rate from our LDA-based approach was also relatively low compared to other techniques, which proves that this approach produced better results consistently.

6. FUTURE WORK

Our system only constrains the tracker when words and gestures occur simultaneously. However, studies have shown that utterances and gestures forming the same idea unit can occur sequentially as well. [13] According to psychology literature, a person articulates a gesture over a set of words described as a gesture phrase[11]. We can extend our work to add constraints when a key set of words, i.e. gesture phrase are spoken instead of just single words.

The process of segmenting the gestures into "gestemes" during gesture recognition simply models the features in a Gaussian Mixture Model using an EM algorithm. While this is an automatic segmentation, the process is highly prone to random initialisation of the EM algorithm. A further analysis of clustering of the pose features is required.

7. CONCLUSIONS

In this paper, we have presented a body pose tracking system that makes use of the context via spoken words to improve tracking in a timely fashion. In other words, the tracking is improved just at the time when the tracking information is important for gesture recognition or deictic resolution. While it might seem that improving deictic resolution using spoken keywords is a limited problem, we believe this idea can be extended to improve the recognition or tracking of more complex gestures. For example, in a lecture scenario, where a lecturer gesticulates to explain a diagram on the board. Such gestures are more natural and more complicated, and the vocabulary of words associated with the same gestures are more extensive.

In the later half of this paper, we used a combination of well-known approaches to improve gesture recognition. We have shown a simple transformation on the pose features can decrease the error rate significantly.

8. ACKNOWLEDGMENTS

We would like to thank Eugene Weinstein for all his dedicated help in setting up the Galaxy system for us, and Ou Wanmei for kindly reviewing our paper and providing good suggestions.

	Pose Features f_n	LDA Features f_n'	PCA	kPCA	Velocity Features ν_n	Velocity and Pose Features ψ_n
Avg Error Rate, $\mu(\%)$	12.16	7.98	26.73	25.09	17.6	13.81
Std Error Rate, $\sigma(\%)$	2.05	1.68	1.84	3.6	1.82	2.86

Table 2: Statistics of Error Rates of various techniques or feature types using a constant feature length of 4. The standard deviation and the average error rate of the LDA features is smaller than the other feature types. This shows that LDA features will lower the error rate consistently

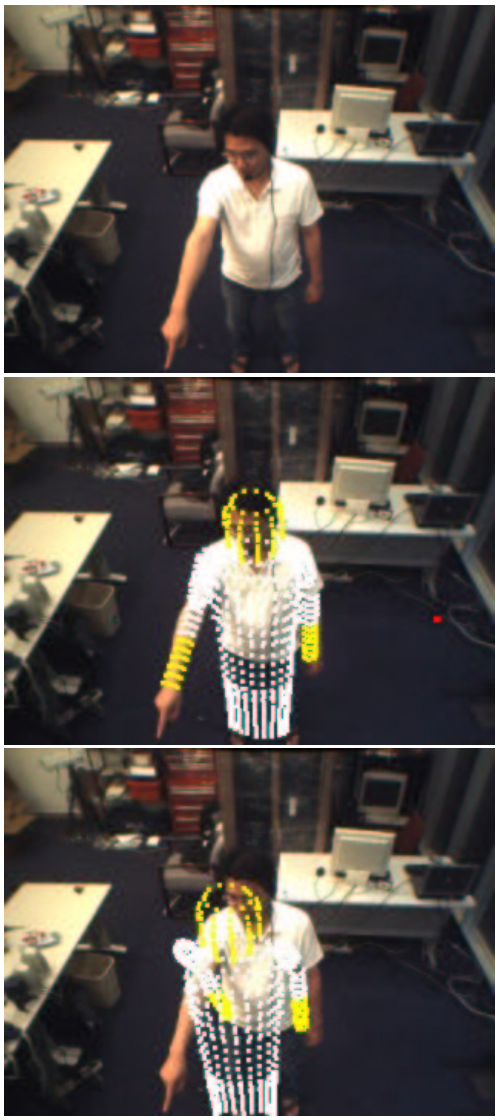

Figure 6: Original image (top) and tracking results from the audio-visual tracker (middle) and the standard vision-based (bottom) algorithms corresponding to frame 1860 of the sequence plotted in Figure 5. Around this frame, the user pronounced the word *'move'*, which induced a *pointing context* (right arm pointing, torso facing the projection screen).

Figure 7: Images of a user articulating 4 different gestures

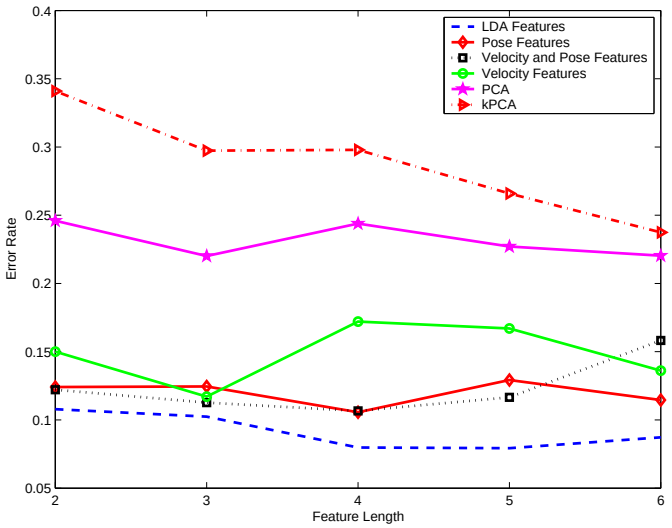

Figure 8: Error rates of using different types of features vs their feature length. The graph shows that the LDA-based approach produced lower error rates consistently over different feature lengths

9. REFERENCES

[1] R. A. Bolt. Put that there: Voice and gesture at the graphics interface. In *7th annual conf. on Computer Graphics and Interactive Techniques*, pages 262–270. ACM Press, 1980.

[2] A. Corradini, R. Wesson, and P. R. Cohen. A map-based system using speech and 3d gestures for pervasive computing. In *Int'l Conf. on Multimodal Interfaces*, 2002.

[3] Q. Delamarre and O. D. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *Proceedings of ICCV'99*, pages 716–721, 1999.

[4] D. Demirdjian and T. Darrell. 3-d articulated pose tracking for untethered deictic reference. In *Int'l Conf. on Multimodal Interfaces*, 2002.

[5] H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *Int'l Conf. on Multimodal Interfaces*, pages 175–182, 2004.

[6] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing*. Prentice Hall, Upper Sadle River, New Jersey, 2001.

[7] C. Hundtofte, G. Hager, and A. Okamura. Building a task language for segmentation and recognition of user input to cooperative manipulation systems. In *IEEE Virtual Reality Conference (HAPTICS 2002)*, Orlando, Florida, 2002.

[8] N. Jojic, M. Turk, and T. Huang. Tracking articulated objects in dense disparity maps. In *International Conference on Computer Vision*, pages 123–130, 1999.

[9] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, and X. Li. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Int'l Conf. on Multimodal Interfaces*, Vancouver, B. C., Canada, 2003.

[10] D. B. Koons, C. J. Sparrell, and K. R. Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*, M. T. Maybury, 1993. AAAI Press/MIT Press Cambridge, MA.

[11] D. McNeill. *Hand and Mind*. The University Chicago Press, Chicago IL, 1992.

[12] S. L. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Int'l conf. of Computer Human Interfaces(CHI 99)*, Pittsburgh, Pennsylvania, 1999.

[13] S. L. Oviatt. Ten myths of multimodal interaction. In *CACM*, volume 42(11), pages 74–81, 1999.

[14] S. L. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *CHI*, pages 415–422, 1997.

[15] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, pages 257–286, 2002.

[16] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue. Galaxy-2: A reference architecture for conversational system development. In *Int'l conf. of Speech and Language Processing*, Sydney, Australia, 1998.

[17] R. Sharma, J. Cai, S. Chakravathy, I. Poddar, and Y. Sethi. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. In *Int'l Conf. on Face and Gesture Recognition*, Grenoble France, 2000.

[18] R. Stiefelhagen, C. Fugen, H. H. P. Gieselmann, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, gaze and gestures. In *Int'l Conf. on Intelligent Robots and Systems*, Sendai, Japan, 2004.