M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 466
To appear IEEE PAMI '98; Submitted 4/26/96

Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video

Thad Starner, Joshua Weaver, and Alex Pentland

Room E15-383, The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge MA 02139
thad,joshw,sandy@media.mit.edu

Abstract

We present two real-time hidden Markov model-based systems for recognizing sentence-level continuous American Sign Language (ASL) using a single camera to track the user's unadorned hands. The first system observes the user from a desk mounted camera and achieves 92% word accuracy. The second system mounts the camera in a cap worn by the user and achieves 98% accuracy (97% with an unrestricted grammar). Both experiments use a 40 word lexicon.

Categories: Gesture Recognition, Hidden Markov Models, Wearable Computers, Sign Language, Motion and Pattern Analysis.

1 Introduction

While there are many different types of gestures, the most structured sets belong to the sign languages. In sign language, each gesture already has assigned meaning, and strong rules of context and grammar may be applied to make recognition tractable. American Sign Language (ASL) is the language of choice for most deaf in the United States. ASL uses approximately 6000 gestures for common words and finger spelling for communicating obscure words or proper nouns. However, the majority of signing is with full words, allowing signed conversations to proceed at about the pace of spoken conversation. ASL's grammar allows more flexibility in word order than English and sometimes uses redundancy for emphasis. Another variant, Signed Exact English (SEE), has more in common with spoken English but is not as widespread in America.

Conversants in ASL may describe a person, place, or thing and then point to a place in space to store that object temporarily for later reference [14]. For the purposes of this experiment, this aspect of ASL will be ignored. Furthermore, in ASL the eyebrows are raised for a question, relaxed for a statement, and furrowed for a directive. While we have also built systems that track facial features [4, 9], this source of information will not be used to aid recognition in the task addressed here.

1.1 Related Work

Following a similar path to early speech recognition, many previous attempts at machine sign language recognition concentrate on isolated signs or fingerspelling. Space does not permit a thorough review [19], but, in general, most attempts either relied on instrumented gloves or a desktop-based camera system and used a form of template matching or neural nets for recognition. However, current extensible systems are beginning to employ hidden Markov models (HMM's).

Hidden Markov models are used prominently and successfully in speech recognition and, more recently, in handwriting recognition. Consequently, they seem ideal for visual recognition of complex, structured hand gestures as are found in sign languages. Explicit segmentation on the word level is not necessary for either training or recognition. Language and context models can be applied on several different levels, and much related development of this technology has been done by the speech recognition community [6].

When the authors first reported this project in 1995 [15, 18], very few uses of HMM's were found in the computer vision literature [22, 13]. At the time, continuous density HMM's were beginning to appear in the speech community; continuous gesture recognition was scarce; gesture lexicons were very small; and automatic training through Baum-Welch re-estimation was uncommon. Results were not reported with the standard accuracy measures accepted in the speech and handwriting recognition communities, and training and testing databases were often identical or dependent in some manner.

Since this time, HMM-based gesture recognizers for other tasks have appeared in the literature [21, 2], and, last year, several HMM-based continuous sign language systems were demonstrated. In a submission to UIST'97, Liang and Ouhyoung's work in Taiwanese Sign Language [8] shows very encouraging results with a glove-based recognizer. This HMM-based system recognizes 51 postures, 8 orientations, and 8 motion primitives. When combined, these constituents can form a lexicon of 250 words which can be continuously recognized in real-time with 90.5% accuracy. At ICCV'98, Vogler and Metaxas described a desk-based 3D camera system that achieves 89.9% word accuracy on a 53 word lexicon [20]. Since the vision process is computationally expensive in this implementation, an electromagnetic tracker is used interchangeably with the 3 mutually orthogonal calibrated cameras for collecting experimental data.

1.2 The Task

In this paper, we describe two extensible systems which use one color camera to track unadorned hands in real time and interpret American Sign Language using hidden Markov models. The tracking stage of the system does not attempt a fine description of hand shape, instead concentrating on the evolution of the gesture through time. Studies of human sign readers suggest that surprisingly little hand detail is necessary for humans to interpret sign language [10, 14]. In fact, in movies shot from the waist up of isolated signs, Sperling *et al.* [14] show that the movies retain 85% of their full resolution intelligibility when subsampled to 24 by 16 pixels! For this experiment, the tracking process produces only a coarse description of hand shape, orientation, and trajectory. The resulting information is input to a HMM for recognition of the signed words.

While the scope of this work is not to create a user independent, full lexicon system for recognizing ASL, the system is extensible toward this goal. The "continuous" sign language recognition of full sentences demonstrates the feasibility of recognizing complicated series of gestures. In addition, the real-time recognition techniques described here allow easier experimentation, demonstrate the possibility of a commercial product in the future, and simplify archival of test data. For this recognition system, sentences of the form "personal pronoun, verb, noun, adjective, (the same) personal pronoun" are to be recognized. This structure allows a large variety of meaningful sentences to be generated using randomly chosen words from each class as shown in Table 1. Six personal pronouns, nine verbs, twenty nouns, and five adjectives are included for a total lexicon of forty words. The words were chosen by paging through Humphries *et al.* [7] and selecting those words which would generate coherent sentences given the grammar constraint. Words were not chosen based on distinctiveness or lack of detail in the finger positioning. Note that finger position plays an important role in several of the signs (pack vs. car, food vs. pill, red vs. mouse, etc.)

Table 1: ASL Test Lexicon

part of speech	*vocabulary*
pronoun	I, you, he, we, you(pl), they
verb	want, like, lose, dontwant, dontlike, love, pack, hit, loan
noun	box, car, book, table, paper, pants, bicycle, bottle, can, wristwatch, umbrella, coat, pencil, shoes, food, magazine, fish, mouse, pill, bowl
adjective	red, brown, black, gray, yellow

2 Hidden Markov Modeling

Due to space limitations, the reader is encouraged to refer to the existing literature on HMM evaluation, estimation, and decoding [1, 6, 11, 23]. A tutorial relating HMM's to sign language recognition is provided in the first author's Master's thesis [15].

The initial topology for an HMM can be determined by estimating how many different states are involved in specifying a sign. Fine tuning this topology can be performed empirically. In this case, an initial topology of 5 states was considered sufficient for the most complex sign. To handle less complicated signs, skip transitions were specified which allowed the topology to emulate a strictly 3 or 4 state HMM. While different topologies can be specified per sign explicitly, the above method allows training to adapt the HMM automatically without human intervention. However, after testing several different topologies, a four state HMM with one skip transition was determined to be appropriate for this task (Figure 1).

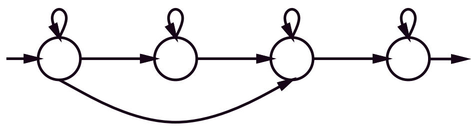

Figure 1: The four state HMM used for recognition.

3 Feature extraction and hand ambiguity

Previous systems have shown that, given strong constraints on viewing, relatively detailed models of the hands can be recovered from video images [3, 12]. However, many of these constraints conflict with recognizing ASL in a natural context, since they either require simple, unchanging backgrounds (unlike clothing); do not allow occlusion; require carefully labelled gloves; or are difficult to run in real time.

In this project, we track the hands using a single camera in real-time without the aid of gloves or markings. Only the natural color of the hands is needed. For vision-based sign recognition, the two possible mounting locations for the camera are in the position of an observer of the signer or from the point of view of the signer himself. These two views can be thought of as second-person and first-person viewpoints, respectively.

Training for a second-person viewpoint is appropriate in the rare instance when the translation system is to be worn by a hearing person to translate the signs of a mute or deaf individual. However, such a system is also appropriate when a signer wishes to control or dictate to a desktop computer as is the case in the first experiment. Figure 2 demonstrates the viewpoint of the desk-based experiment.

Figure 2: View from the desk-based tracking camera. Images are analyzed at 320x240 resolution.

The first-person system observes the signer's hands from much the same viewpoint as the signer himself. Figure 3 shows the placement of the camera in the cap used in the second experiment, and demonstrates the resulting viewpoint. The camera was attached to an SGI for development; however, current hardware allows for the entire system to be unobtrusively embedded in the cap itself as a wearable computer. A matchstick-sized camera such as the Elmo QN401E can be embedded in front seam above the brim. The brim can be made into a relatively good quality speaker by lining it with a PVDF transducer (used in thin consumer-grade stereo speakers). Finally a PC/104-based CPU, digitizer, and batteries can be placed at the back of the head. See Starner *et al.* [17] and the MIT Wearable Computing Site (http://wearables.www.media.mit.edu/projects/wearables/) for more detailed information about wearable computing and related technologies.

Figure 3: The hat-mounted camera, pointed downward towards the hands, and the corresponding view.

A wearable computer system provides the greatest utility for an ASL to spoken English translator. It can be worn by the signer whenever communication with a non-signer might be necessary, such as for business or on vacation. Providing the signer with a self-contained and unobtrusive first-person view translation system is more feasible than trying to provide second-person translation systems for everyone whom the signer might encounter during the day.

For both systems, color NTSC composite video is captured and analyzed at 320 by 243 pixel resolution. This lower resolution avoids video interlace effects. A Silicon Graphics 200MHz R4400 Indy workstation maintains hand tracking at 10 frames per second, a frame rate which Sperling *et al.* [14] found sufficient for human recognition. To segment each hand initially, the algorithm scans the image until it finds a pixel of the appropriate color, determined by an *a priori* model of skin color. Given this pixel as a seed, the region is grown by checking the eight nearest neighbors for the appropriate color. Each pixel checked is considered part of the hand. This, in effect, performs a simple morphological dilation upon the resultant image that helps to prevent edge and lighting aberrations. The centroid is calculated as a by-product of the growing step and is stored as the seed for the next frame. Since the hands have the same skin tone, the labels "left hand" and "right hand" are simply assigned to whichever blob is leftmost and rightmost.

Note that an *a priori* model of skin color may not be appropriate in some situations. For example, with a mobile system, lighting can change the appearance of the hands drastically. However, the image in Figure 3 provides a clue to addressing this problem, at least for the first-person view. The smudge on the bottom of the image is actually the signer's nose. Since the camera is mounted on a cap, the nose always stays in the same place relative to the image. Thus, the signer's nose can be used as a calibration object for generating a model of the hands' skin color for tracking. While this calibration system has been prototyped, it was not used in these experiments.

After extracting the hand blobs from the scene, second moment analysis is performed on each blob. A sixteen element feature vector is constructed from each hand's x and y position, change in x and y between frames, area (in pixels), angle of axis of least inertia (found by the first eigenvector of the blob) [5], length of this eigenvector, and eccentricity of bounding ellipse.

When tracking skin tones, the above analysis helps to model situations of hand ambiguity implicitly. When a hand occludes either the other hand or the face (or the nose in the case of the wearable version), color tracking alone can not resolve the ambiguity. Since the face remains in the same area of the frame, its position can be determined and discounted. However, the hands move rapidly and occlude each other often. When occlusion occurs, the hands appear to the above system as a single blob of larger than normal area with significantly different moments than either of the two hands in the previous frame. In this implementation, each of the two hands is assigned the features of this single blob whenever occlusion occurs. While not as informative as tracking each hand separately, this method still retains a surprising amount of discriminating information. The occlusion event itself is implicitly modeled, and the combined position and moment information are retained. This method, combined with the time context provided by hidden Markov models, is sufficient to distinguish between many different signs where hand occlusion occurs.

4 The second person view: a desk-based recognizer

The first experimental situation explored was the second person view: a desk-based recognizer. In this experiment 500 sentences were obtained, but 22 sentences were eliminated due to subject error or outlier signs. In general, each sign is 1 to 3 seconds long. No intentional pauses exist between signs within a sentence, but the sentences themselves are distinct. For testing purposes, 384 sentences were used for training, and 94 were reserved for testing. The test sentences are not used in any portion of the training process.

For training, the sentences are divided automatically in five equal portions to provide an initial segmention into component signs. Then, initial estimates for the means and variances of the output probabilities are provided by iteratively using Viterbi alignment on the training data and then recomputing the means and variances by pooling the vectors in each segment. Entropic's Hidden Markov Model ToolKit (HTK) is used as a basis for this step and all other HMM modeling and training tasks. The results from the initial alignment program are fed into a Baum-Welch re-estimator, whose estimates are, in turn, refined in embedded training which ignores any initial segmentation. For recognition, HTK's Viterbi recognizer is used both with and without the part-of-speech grammar based on the known form of the sentences. Contexts are not used since they would require significantly more data to train. However, a similar effect can be achieved with the strong grammar in this data set. Recognition occurs five times faster than real time.

Word recognition accuracy results are shown in Table 2; when different, the percentage of words correctly recognized is shown in parentheses next to the accuracy rates. Accuracy is calculated by

$$Acc = \frac{N - D - S - I}{N}$$

where N is the total number of words in the test set, D is the number of deletions, S is the number of substitutions, and I is the number of insertions. Note that, since all errors are accounted against the accuracy rate, it is possible to get large negative accuracies (and corresponding error rates of over 100%). When using the part-of-speech grammar (pronoun, verb, noun, adjective, pronoun), insertion and deletion errors are not possible since the number and class of words allowed is known. Thus, all errors are vocabulary substitutions when this grammar is used (and accuracy is equivalent to percent correct). Assuming independence, random chance would result in a percent correct of 13.9%, calculated by averaging over the likelihood of each part-of-speech being correct. Without the grammar, the recognizer is allowed to match the observation vectors with any number of the 40 vocabulary words in any order. In fact, the number of words produced by the recognizer can be up to the number of samples in the sentence! Thus, deletion (D), insertion (I), and substitution (S) errors are possible in the "unrestricted grammar" tests, and a comparison to random chance becomes irrelevant. The absolute number of errors of each type are listed in Table 2. Many of the insertion errors correspond to signs with repetitive motion.

An additional "relative features" test is provided in the results. For this test, absolute (x, y) position is removed from the feature vector. This provides a sense of how the recognizer performs when only relative features are available. Such may be the case in daily use; the signer may not place himself in the same location each time the system is used.

Table 2: Word accuracy of desk-based system

experiment	*training set*	*independent test set*
all features	94.1%	91.9%
relative features	89.6%	87.2%
all features & unrestricted grammar	81.0% (87%) (D=31, S=287, I=137, N=2390)	74.5% (83%) (D=3, S=76, I=41, N=470)

Word accuracies; percent correct in parentheses where different. The first test uses the strong part-of-speech grammar and all feature elements. The second test removes absolute position from the feature vector. The last test again uses all features but only requires that the hypothesized output be composed of words from the lexicon. Any word can occur at any time and any number of times.

The 94.1% and 91.9% accuracies using the part-of-speech grammar show that the HMM topologies are sound and that the models generalize well. However, the subject's variability in body rotation and position is known to be a problem with this data set. Thus, signs that are distinguished by the hands' positions in relation to the body were confused since the absolute positions of the hands in screen coordinates were measured. With the relative feature set, the absolute positions of the hands are be removed from the feature vector. While this change causes the error rate to increase slightly, it demonstrates the feasibility of allowing the subject to vary his location in the room while signing, possibly removing a constraint from the system.

The error rates of the "unrestricted" experiment better indicate where problems may occur when extending the system. Without the grammar, signs with repetitive or long gestures were often inserted twice for each actual occurrence. In fact, insertions caused more errors than substitutions. Thus, the sign "shoes" might be recognized as "shoes shoes," which is a viable hypothesis without a language model. However, a practical solution to this problem is to use context training and a statistical grammar.

5 The first person view: a wearable-based recognizer

For the second experiment, the same 500 sentences were collected by a different subject. Sentences were resigned whenever a mistake was made. The full 500 sentence database is available from anonymous ftp at whitechapel.media.mit.edu under pub/asl. The subject took care to look forward while signing so as not to confound the tracking with head rotation, though variations can be seen. Often, several frames at the beginning and ending of a sentence's data contain the hands at a resting position. To take this in account, another token, "silence" (in deference to the speech convention), was added to the lexicon. While this "sign" is trained with the rest, it is not included when calculating the accuracy measurement.

The resulting word accuracies from the experiment are listed in Table 3. In this experiment 400 sentences were used for training, and an independent 100 sentences were used for testing. A new grammar was added for this experiment. This grammar simply restricts the recognizer to five word sentences without regard to part of speech. Thus, the percent correct words expected by chance using this "5-word" grammar would be 2.5%. Deletions and insertions are possible with this grammar since a repeated word can be thought of as a deletion and an insertion instead of two substitutions.

Table 3: Word accuracy of wearable computer system

grammar	*training set*	*independent test set*
part-of-speech	99.3%	97.8%
5-word sentence	98.2% (98.4%) (D = 5, S=36, I=5 N =2500)	97.8%
unrestricted	96.4% (97.8%) (D=24, S=32, I=35, N=2500)	96.8% (98.0%) (D=4, S=6, I=6, N=500)

Word accuracies; percent correct in parentheses where different. The 5-word grammar limits the recognizer output to 5 words selected from the vocabulary. The other grammars are as before.

Interestingly, for the part-of-speech, 5-word, and unrestricted tests, the accuracies are essentially the same, suggesting that all the signs in the lexicon can be distinguished from each other using this feature set and method. As in the previous experiment, repeated words represent 25% of the errors in the unrestricted grammar test. In fact, if a simple repeated word filter is applied post process to the recognition, the unrestricted grammar test accuracy becomes 97.6%, almost exactly that of the most restrictive grammar! Looking carefully at the details of the part-of-speech and 5-word grammar tests indicate that the same beginning and ending pronoun restriction may have hurt the performance of the part-of-speech grammar! Thus, the strong grammars are superfluous for this task. In addition, the very similar results between fair-test and test-on-training cases indicate that the HMM's training converged and generalized extremely well for the task.

The main result is the high accuracies themselves, which indicate that harder tasks should be attempted. However, why is the wearable system so much more accurate than the desk system? There are several possible factors. First, the wearable system has less occlusion problems, both with the face and between the hands. Second, the wearable data set did not have the problem with body rotation that the first data set experienced. Third, each data set was created and verified by separate subjects, with successively better data recording methods. Controlling for these various factors requires a new experiment, described in the next section.

6 Discussion and Future Work

We have shown a high accuracy computer vision-based method of recognizing sentence-level American Sign Language selected from a 40 word lexicon. The first experiment shows how the system can be used to communicate with a desk-based computer. The second experiment demonstrates how a wearable computer might use this method as part of an ASL to English translator. Both experiments argue that HMM's will be a powerful method for sign language recognition, much as they have been for speech and handwriting recognition. In addition, the experiments suggest that the first person view provides a valid perspective for creating a wearable ASL translator.

While it can be argued that sign evolved to have maximum intelligibility from a frontal view, further thought reveals that sign also may have to be distinguishable by the signer himself, both for learning and to provide control feedback. To determine which view is superior for recognition, we have begun a new experiment. Native signers will be given a task to complete. The task will be designed to encourage a small vocabulary (*e.g.* a few hundred words) and to encourage natural sign. Four views of the signers will be recorded simulaneously: a stereo pair from the front, a view from the side, and the wearable computer view. Thus, both 3D and 2D tracking from various views can be compared directly.

Head motion and facial gestures also have roles in sign which the wearable system would seem to have trouble addressing. In fact, uncompensated head rotation would significantly impair the current system. However, as shown by the effects in the first experiment, body/head rotation is an issue from either viewpoint. Simple fiducials, such as a belt buckle or lettering on a t-shirt may be used to compensate tracking or even provide additional features. Another option for the wearable system is to add inertial sensors to compensate for head motion. In addition, EMG's may be placed in the cap's head band along the forehead to analyze eyebrow motion as has been discussed by Picard [9]. In this way facial gesture information may be recovered.

As the system grows in lexicon size, finger and palm tracking information may be added. This may be as simple as counting how many fingers are visible along the contour of the hand and whether the palm is facing up or down. In addition, tri-sign context models and statistical grammars may be added which may reduce error up to a factor of eight if speech and handwriting trends hold true for sign [16].

These improvements do not address user independence. Just as in speech, making a system which can understand different subjects with their own variations of language involves collecting data from many subjects. Until such a system is tried, it is hard to estimate the number of subjects and the amount of data that would comprise a suitable training database. Independent recognition often places new requirements on the feature set as well. While the modifications mentioned above may be initially sufficient, the development process is highly empirical.

Similarly, we have not yet addressed the problem of finger spelling. Changes to the feature vector to address finger information are vital, but adjusting the context modeling is also of importance. With finger spelling, a closer parallel can be made to speech recognition. Tri-sign context occurs at the sub-word level while grammar modeling occurs at the word level. However, this is at odds with context across word signs. Can tri-sign context be used across finger spelling and signing? Is it beneficial to switch to a separate mode for finger spelling recognition? Can natural language techniques be applied, and if so, can they also be used to address the spatial positioning issues in ASL? The answers to these questions may be key to creating an unconstrained sign language recognition system.

7 Acknowledgements

The authors would like to thank Tavenner Hall for her help editing this document. This work is supported by BT and the Things That Think consortium at the MIT Media Laboratory.

8 References

[1] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[2] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick, and A. Pentland. Invariant features for 3-d gesture recognition. In *Second Intl. Conf. on Face and Gesture Recogn.*, pages 157–162, 1996.

[3] B. Dorner. Hand shape identification and tracking for sign language interpretation. In *IJCAI Workshop on Looking at People*, 1993.

[4] I. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, Texas, Nov. 1994.

[5] B. Horn. *Robot Vision.* MIT Press, Cambridge, MA, 1986.

[6] X.D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[7] T. Humphries, C. Padden, and T. O'Rourke. *A Basic Course in American Sign Language.* T. J. Publ., Inc., Silver Spring, MD, 1990.

[8] R. Liang and M. Ouhyoung. A real-time continuous gesture interface for Taiwanese Sign Language. In *Submitted to UIST*, 1997.

[9] R. Picard. Toward agents that recognize emotion. In *Imagina98*, 1998.

[10] H. Poizner, U. Bellugi, and V. Lutes-Driscoll. Perception of American Sign Language in dynamic point-light displays. *J. Exp. Pyschol.: Human Perform.*, 7:430–440, 1981.

[11] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

[12] J. M. Rehg and T. Kanade. DigitEyes: vision-based human hand tracking. School of Computer Science Technical Report CMU-CS-93-220, Carnegie Mellon University, December 1993.

[13] J. Schlenzig, E. Hunter, and R. Jain. Recursive identification of gesture inputs using hidden Markov models. *Proc. Second Annual Conference on Applications of Computer Vision*, pages 187–194, December 1994.

[14] G. Sperling, M. Landy, Y. Cohen, and M. Pavel. Intelligible encoding of ASL image sequences at extremely low information rates. *Comp. Vis., Graph., and Img. Proc.*, 31:335–391, 1985.

[15] T. Starner. Visual recognition of American Sign Language using hidden Markov models. Master's thesis, MIT, Media Laboratory, February 1995.

[16] T. Starner, J. Makhoul, R. Schwartz, and G. Chou. On-line cursive handwriting recognition using speech recognition methods. In *ICASSP*, pages 125–128, 1994.

[17] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, and A. Pentland. Augmented reality through wearable computing. *Presence*, 6(4):386–398, Winter 1997.

[18] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using hidden Markov models. Technical Report 375, MIT Media Lab, Perceptual Computing Group, 1995. Earlier version appeared ISCV'95.

[19] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desktop and wearable computer based video. Technical Report 466, Perceptual Computing, MIT Media Laboratory, July 1998.

[20] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *ICCV*, Bombay, 1998.

[21] A. D. Wilson and A. F. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, Florida, November 1995.

[22] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov models. *Proc. Comp. Vis. and Pattern Rec.*, pages 379–385, 1992.

[23] S. Young. *HTK: Hidden Markov Model Toolkit V1.5.* Cambridge Univ. Eng. Dept. Speech Group and Entropic Research Lab. Inc., Washington DC, 1993.