ROBUST GESTURE RECOGNITION FOR AUGMENTATIVE COMMUNICATION USING A DYNAMIC BINARY FRAME OF REFERENCE

Jesus De La Rosa Estanol [13], Deryk Beal [2], Eric Bouffet [12], Brian Kavanagh [12], Tom Chau [13]
[1] University of Toronto, [2] The Hospital for Sick Children, [3] Bloorview Macmillan Children's Center

***Abstract*— A key challenge in gesture recognition for augmentative communication is the ability to characterize a given gesture by its fundamental and invariant properties. This paper proposes a new approach to this problem, based on characterizing gestures with primitive labels such as "left-up to right-down". It uses movement as a recognition cue and the relative positions between key motion descriptors to characterize the gesture. This approach is robust to rotation and changes in scale and is performer independent. Further, the method can tolerate cluttered backgrounds and does not require the user to wear any accessories. A single representative example is sufficient to characterize a given gesture and no training is required. Experiments with a set of gestures performed by different individuals in a cluttered environment demonstrate the robustness of the approach. Implications for gesture-based augmentative communication are briefly discussed.**

***Index Terms*— gesture recognition, motion detection, human computer interface, augmentative communication.**

I. INTRODUCTION

THE inability to communicate places a large amount of stress, not only on patients but also on families and caregivers. Families and caregivers are often unable to interpret the needs, preferences and feelings of patients [1]. To address this communication need, a system is required to compensate for the lack of communication ability. Typical solutions may take the form of printed material (e.g. a picture board) or an electronic device. These alternative communication strategies form a field of study known as Augmented and Alternative Communication (AAC) [2].

Gesture recognition (GR) systems gather information to determine limb or body position to allow a human-machine interaction. This interaction can be contact or non contact. Examples of GR systems are found throughout the literature. Azoz et al. presented a system to recognize arm movements [3]. Skin color was used as a cue to perform head and hand localization. Their work was based on the mathematical modeling of the arm (elbow-shoulder-hand) and the relation between their parts, thus defining implicit constraints that made it an accurate but rigid model. Psarrou et al. presented a system capable of gesture recognition which also described movement patterns in a typical office [4]. However, they relied heavily on statistical data to account for the characteristics and variability of the event's duration. Nishimura and Mukai performed similar experiments, employing low resolution images to extract features and to counter two common problems in GR, namely threshold adjustment and offline training [5]. Starner et al. at MIT developed a real time system for American Sign Language (ASL) using Hidden Markov Models. They relied on motion detection, orientation and hands/arms trajectories as well as cueing with the skin's natural color. Unfortunately, their system only worked with a single user [6]. Matsugu et al. used a convolutional neural network for GR. They determined face location using color cues and isolated the position of the eyes and mouth. They were able to achieve subject-independent recognition and could determine three gestures: smiling, laughing and a neutral face. Due to their neural network approach, at least 20,000 images were required [7].

In this paper, we propose the Dynamic Binary Frame of Reference (DBFR) method for gesture recognition. This method has several advantages over existing techniques including tolerance to different degrees of rotation, scale and geometry invariance of the moving object and performer independence. This method does not require a rigid mathematical model. Thus, it is capable of recognizing the same gesture performed using different body parts, such as an arm, hand or finger. The system operates in cluttered backgrounds and does not impose restrictions on the user by way of accessories or sensors. The main feature of the system is its capability to operate without any adjustment, assuming that the gesture performed is defined in the gesture database. Additionally, gestures can be linguistically described using combinations of lay terms such as up, right, left and down.

II. METHOD DESCRIPTION

The Dynamic Binary Frame of Reference method uses a loose frame of reference to determine the direction of movement at a given time. Its binary nature arises from its consideration of two possible motion descriptors per axis, as described later in the paper.

The GR system records video sequences of varying length at a fixed frame rate and recognizes the gestures performed. The overall GR algorithm is described as follows:

1) Image acquisition
2) Image undersampling
3) Calculation of first difference.
4) Dilatation.
5) Calculation of second difference.
6) Cumulative subtraction.

7) Local maxima extraction.
8) Relative direction.

Images taken from a webcam typically have a resolution of 320×240 pixels. Thus, processing such images becomes a costly computational task. We acquire images using a resolution of 160×120 and then the images are undersampled to 16×12 pixels. Assume that we have acquired and downsampled a video sequence of n frames. Each image is then subtracted from its preceeding one in a serial fashion providing an *apparent motion* matrix M_{am} for each image pair, as shown in equation (1). The resulting video sequence contains only $n-1$ frames.

$$M_{am}(n-1) = I(n) - I(n-1) \tag{1}$$

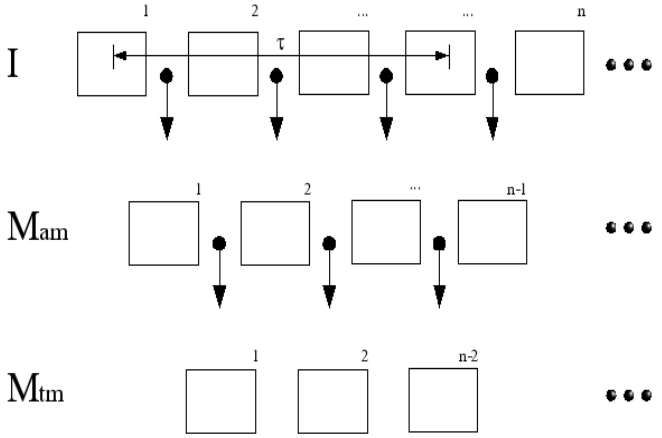

Fig. 1. Original sequence I; first difference (apparent motion)M_{am}; second difference (true motion)M_{tm}

Some recognition algorithms stop at this substraction step, but the information contained in M_{am} is not completely true due to movement and natural hardware noise inherent in acquired images. These noise variations continuously produce changes between frames even if scene or lighting conditions remain constant. To overcome this problem, it is imperative to validate the information in our matrix through an additional step. To this end, we compute *true movement*, that is, continuous and uninterrupted movement within a sequence of frames, without any assumptions about the instantaneous direction of movement. By expanding or dilating apparent motion frames with a binary filter, f, and performing a logical *AND* on $M_{ex}(n-1)$ and $M_{ex}(n)$ we can accomodate future directions of movement.

$$M_{ex}(n-1) = M_{am}(n-1) \times f \tag{2}$$

where:

$$f = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{3}$$

M_{ex} represents the expanded matrix and $\times$ symbolizes the correlation operation.

Next, we compute the difference between successive expanded matrices to obtain a true motion matrix, M_{tm}.

$$M_{tm}(n-2) = M_{ex}(n-1) - M_{ex}(n-2) \tag{4}$$

After this operation, the number of useful frames is $n-2$.

We perform a second subtraction of consecutive *true motion* matrices M_{tm}.

$$T(\tau) = \sum_{n}^{\tau} M_{tm}(n) - M_{tm}(n-1) \tag{5}$$

where, the entry on the i^{th} row and j^{th} column is

$$T_{ij} = \begin{cases} -1 & \text{Initial position} \\ 0 & \text{Partial or no movement} \\ 1 & \text{Final position} \end{cases} \tag{6}$$

This step provides a new set of matrices, T, with entries $\{C_i, C_n, C_f\} = \{-1, 0, 1\}$, where C_i represents the start of the motion, C_n stands for partial or null movement and C_f indicates the end of the trajectory. The subtraction of two true motion matrices will result in partial trajectories. By performing an addition of all the partial results over a period τ we arrive at the true trajectory of the object.

An *accumulative movement* matrix M_c given by equation (7) can be computed.

$$M_c(n-2) = \sum_{n-2}^{\tau} M_{tm}(n-2) \tag{7}$$

where τ_i is a sequence of i consecutive frames.

In M_c, we determine the elements with maximal movement denoted by C_m as exemplified in (8).

$$M_c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 2 & 2 & 2 & 2 \\ 3 & 5 & 4 & 3 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 2 & 2 & 2 & 2 \\ 3 & C_m & 4 & 3 \end{bmatrix} \tag{8}$$

We construct a table containing motion descriptors given by their relative positions between elements indicated by five motion descriptors: *left (L), right (R), up (U), down (D), do not care (X)*. The "do not care" (X) symbol is included to deal with uncertainty during certain stages of movement. For example, on a vertical trace, it is clear that the object is moving upward but it might also contain a slight lateral movement. To avoid the use of thresholds, these ill-defined traces are discarded.

The matrix (9) shows an example of C_m, C_i, C_f. The associated motion descriptors are listed in Table I.

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ C_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & C_m & 0 & C_f \end{bmatrix} \tag{9}$$

This matrix provides valuable relative information about the observed gesture, which allows recognition to be invariant to a certain degree of rotation and changes in size or shape.

Additionally, it creates a map of relative positions between initial and final positions for each period τ.

TABLE I

RELATIVE POSITIONS TO C_m. MOTION DESCRIPTORS: LEFT (L), RIGHT (R), UP (U), DOWN (D), DO NOT CARE (X)

	C_i	C_f
C_m	L	R
C_m	D	X

The simple nature of the above motion descriptors permits the creation of gesture templates by hand even without having to acquire image sequences. Therefore, we are able to train the computer about an admissible movement using a single representative example.

III. DESCRIPTION OF THE SYSTEM

A PC Celeron at 800*MHz* is employed to perform processing and recognition. A low cost USB webcam records images at 8 frames per second using a resolution of 160×120 pixels. We do not use any color information to process or recognize the gestures. Hence, the green component of the images is used and it is treated as a gray image.

A single moving object is monitored, where the physical movement of each gesture resembles the pen stroke of its written version. The database of gestures (DBG) currently contains four gestures with the following symbols: “alpha”, “gamma”, “o” and “*N*”. These gestures are performed by an able-bodied user. For each gesture, a representative movement is recorded individually in a cluttered environment, as shown in Figure 2. Another possibility is to manually encode the gestures. For example, a *gamma* gesture would be encoded with the following description: right-down, left-down, left-up, right-up or in motion descriptor codes RD, LD, LU, RU.

Fig. 2. Sequence of frames of a gamma gesture

The system uses a sequence of frames of varying length containing a single gesture as an input. There are no postural constraints imposed on the performer. However, the camera defines a maximum gesture speed due to acquisition time. The system’s output contains the linguistic description of the gesture. Two tests are performed to measure the system’s accuracy and computational load.

IV. RESULTS

Three subjects performed different gestures that are analyzed and recognized according to the DBG. To determine if the motion analyzed is in the DBG a scoring system was employed. Each motion descriptor pair obtained from the video sequence was compared against the DBG and only the entries with equal characteristics scored points. No partial points were given for mismatches or partial matches, therefore a perfect correlation between pairs must be observed. The system currently does not deal with variations or missing motion descriptors. These variations may occur in the video sequence when a gesture is performed at a high speed.

The system was tested with three different subjects. Subjects had no previous training. Each subject performed each gesture three times and in some cases, gestures were performed in different ways, namely, by varying the distance between the performer and the camera; and, by varying the body part employed, for example the shoulder, elbow, wrist or knuckle as shown in Figure 3. The only change between trials was camera positioning. Thus, the system’s capability to process inputs from different body parts was tested.

Fig. 3. Different moving body parts recognized by the system. Images shown are clockwise starting from the top-left: knuckle, right shoulder-elbow, left shoulder-elbow and wrist

The following results were obtained:

TABLE II

RESULTS

	True Positives	Negatives	False positives
Subject 1	66.6%	11.1%	22.2%
Subject 2	77.7%	11.1%	11.1%
Subject 3	100%	0.0%	0.0%

The negative results are related to the limited acquisition speed of the camera and high speed gestures that created empty slots or missing motion descriptors.

A second test was performed to measure the computational load required using different resolutions. Computational load was measured by using a set of different undersampled images. First the original size of 160×120 pixels was used (grid size of 1). The image size was gradually reduced to 4×3 pixels (grid size of 40). As shown in Figure 4, the use of low resolution images exponentialy reduces the computational cost. A grid size between 10 to 20 pixels showed good performance and reasonable processing time. Processing each frame at these resolutions takes on average 0.160 milliseconds or at most 0.8% of the time required to process larger images of 160×120 pixels.

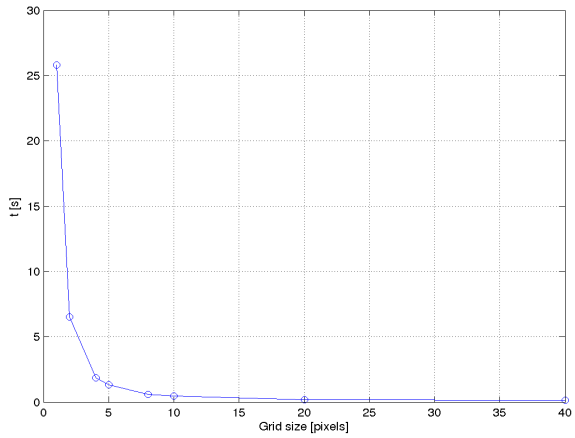

Fig. 4. Average time required to process an "alpha" gesture with different image sizes.

Preliminarily tests showed a potential insensitivity to changes in magnitude or depth of the moving objects in the scene. We also found that when the end trace of a gesture coincides with the start of another, the system is able to recognize each gesture, due to lack of movement between gestures.

V. Conclusion

We presented a novel method to characterize and recognize gestures based in DBFR and its *relative* nature. The presented system shows that recognition of moving objects is not color or shape dependent.

The system characterizes moving objects by their dynamic properties using low resolution features. Additionally, the system robustly copes with different body parts without the need for adjustments. Two further characteristics of the system are worthy of mention. Firstly, the system ignores purely vertical and horizontal movements as their detection requires the definition of ad-hoc thresholds.

Consequently, the system only robustly recognizes angled movements. Secondly, the camera capture speed defines a maximum gesture speed.

The system in clinical settings will provide a human-machine interface to patients. It is expected the system will promote patient communication during periods of voicelessness. The advantage of having a system that requires no adjustments will allow patients to operate the system in a less intrusive and comfortable way. Moreover, the system will not require intensive external assistance.

VI. Future Work

The system is in its early stages. Additional capabilities like the ability to deal with missing motion descriptors as well as the ability to recognize gestures in a continuous stream will be included in the future. As well, we will incorporate the ability to recognize multiple moving objects or areas of interest.

Acknowledgments

The authors would like to thank NSERC, Bloorview MacMillan Children's Center, Bloorview MacMillan Foundation, Hospital for Sick Kids and The University of Toronto.

References

[1] M. B. Happ, *Communication with mechanically ventilated patients: state of the science*, American Association of Critical-Care Nurses: Clinical Issues, Volume 12, Number 2, May 2001

[2] A. M. Cook, S. M. Hussey, *Assistive Technologies, Principles and practices*, Ed. Mosby, 2002

[3] Y. Azoz, L. Devi, R. Sharma,*Tracking Hand Dynamics in Unconstrained Environments*, Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 247-279

[4] A. Psarrou, S. Gong, M. Walter, *Recognition of human gestures and behavior based on motion trajectories*, Image and Vision Computing, Volume 20, Issue 5-6, April 2002, pp. 349-358

[5] T. Nishimura, T. Mukai, S. Nozaki, R. Oka, *Adaptation to gesture performers by an on-line teaching system for spotting recognition of gestures from a time-varying image*, Systems and Computers in Japan, Vol 31, No. 1, 2000, pp. 39-47

[6] T. Starner, J. Weaver, A. Pentland, *Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, December 1998, pp. 1371-1375

[7] M. Matsugu, K. Mori, Y. Mitrari, F. Kaneda, *Subject independent facial expression recognition with robust face detection using a convolutional neural network*, Neural Networks, Vol 16, 2003, pp. 555-559