In Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, pp. 10-15, Nara, Japan

Extracting Gestural Motion Trajectories

Ming-Hsuan Yang and Narendra Ahuja
Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{mhyang, ahuja}@stereo.ai.uiuc.edu

Abstract

This paper is concerned with the extraction of spatio-temporal patterns in video sequences with focus on trajectories of gestural motions associated with American Sign Language. An algorithm is described to extract the motion trajectories of salient features such as human palms from an image sequence. First, motion segmentation of the image sequence is generated based on a multiscale segmentation of the frames and attributed graph matching of regions across frames. This produces region correspondences and their affine transformations. Second, colors of the moving regions are used to determine skin regions. Third, the head and palm regions are identified based on the shape and size of skin regions in motion. Finally, affine transformations defining a region's motion between successive frames are concatenated to construct the region's motion trajectory. Experimental results showing the extracted motion trajectories are presented.

1 Introduction

Gesture recognition plays an important role in the advancement of human computer interaction (HCI) since it provides a natural and efficient way to communicate between humans and computers. With the increasing interest in HCI, there has been rapid growth of studies related to gesture recognition in recent years. Hand gestures comprise the majority of gestures used by humans. In this paper, we will use the term gesture to refer to hand gesture. Most of the work on gesture recognition focuses on static hand gestures or postures [3], [8] with much less attention given to the dynamic characteristics of gestures [11]. Our work is aimed at recognizing gestures of American Sign Language (ASL) using spatio-temporal analysis of gestural motion trajectories.

We perform motion segmentation to group pixels of similar motion into regions and find region correspondences across frames. Although there are many motion regions in each frame, the movements of palm regions contain significant information about the gesture meaning, and therefore palm motion is extracted. Further, ASL experts have pointed out that gesture interpretation requires not only the motion pattern but also relative locations of the hand with respect to other landmark parts of the body such as the head, shoulders, and waist [2]. Therefore, human head region is extracted for use as reference in gesture recognition.

To distinguish among the different moving regions, we use color and geometric characteristics. Both head and palm regions have skin color and have similar elliptic shapes, but differ in size. Motion regions with skin color are first identified, and a connected component analysis is then performed to merge neighboring regions until the shape of the merged region is approximately elliptic. The head and palm regions are discriminated based on difference in size. The palm regions are extracted from each frame and the affine transformations between corresponding regions in successive frames are then computed to obtain motion trajectories. Figure 1 summarizes our algorithm to extract motion trajectories of a gesture from an image sequence.

2 Related Work

Many of the gesture recognition methods utilize specialized colored gloves [5] or markers [8]. Electromagnetic sensor and stereo vision system have also been used to locate a signer in video [14]. Pfinder [4] adopts a maximum a posteriori probability approach to detection and tracking of the human body using simple 2D models for recognizing ASL signs [11]. Motion has also been used for signer localization [6]. However, these approaches have certain restrictions on the signer (e.g. wearing specialized gloves or using markers) or they require that the stationary background have a certain predetermined color.

To overcome the limitations of the individual cues of localization, fusion of cues is explored for head localization in video [7], yet not fully exploited for hand localization as pointed out in a recent review [10]. We describe a method that combines motion, color and geometric analysis for hand localization and motion trajectory computation.

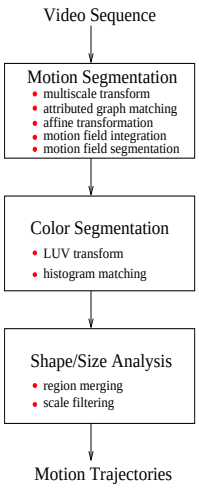

Figure 1. Overview of the algorithm for extracting gestural motion trajectories

This method leads to robustness and accuracy in the extraction of motion trajectories, which in turn helps in recognizing complex hand gestures.

3 Motion Segmentation

In order to capture the dynamic characteristics of hand gestures, we segment an image frame into regions with similar motion. The algorithm processes an image sequence two successive frames at a time. For a pair of frames, (I_t, I_{t+1}), the algorithm identifies regions in each frame comprising the multiscale intraframe structure. Regions at all scales are then matched across frames. Affine transforms re computed for each matched region pair. The affine transform parameters for region at all scales are then used to derive a single motion field which is then segmented to identify the differently moving regions between the two frames. The following sections describe the major steps in the motion segmentation algorithm.

3.1 Multiscale Image Segmentation

Multiscale segmentation is performed using a transform descried in [1] which extracts a hierarchy of regions in each image. The general form of the transform, which maps an image to a family of attraction force fields, is defined by

$$\mathbf{F}(x, y; \sigma_g(x, y), \sigma_s(x, y)) = \int\int_R d_g(\Delta I, \sigma_g(x, y)) \cdot d_s(\vec{r}, \sigma_s(x, y)) \frac{\vec{r}}{||\vec{r}||} dw dv \quad (1)$$

where $R = domain(I(u, v)) \backslash \{(x, y)\}$ and $\vec{r} = (v - x)\vec{i} + (w - y)\vec{j}$. The parameter σ_g denotes a homogeneity scale which reflects the homogeneity of a region to which a pixel belongs and σ_s is spatial scale that controls the neighborhood from which the force on the pixel is computed. The homogeneity of two pixels is given by the Euclidean distance between the associated m-dimensional vectors of pixel values (e.g., $m = 3$ for a color image):

$$\Delta I = |I(x, y) - I(v, w)| \quad (2)$$

The spatial scale parameter, σ_s, controls the spatial distance function, $d_s(\cdot)$, and the homogeneity scale parameter, σ_g, controls the homogeneity distance function, $d_g(\cdot)$. One possible form for these functions satisfying criteria discussed in [13] are unnormalized Gaussian

$$\begin{aligned} d_g(\Delta I, \sigma_g) &\sim \sqrt{2\pi\sigma_g^2} N_{\Delta_I}(0, \sigma_g^2) \\ d_s(\vec{r}, \sigma_s) &\sim \begin{cases} \sqrt{2\pi\sigma_s^2} N_{||\vec{r}||}(0, \sigma_s^2), & ||\vec{r}|| \leq 2\sigma_s \\ 0, & ||\vec{r}|| > 2\sigma_s \end{cases} \end{aligned} \quad (3)$$

The force field encodes the region structure in a manner which allows easy extraction. Region boundaries correspond to diverging force vectors in $\mathbf{F}$ and region skeletons correspond to converging force vectors in $\mathbf{F}$. An increase in σ_g causes less homogeneous structures to be encoded and an increase in σ_s causes large structures to be encoded.

Figure 2 shows an image and its segmented regions at different scales which are displayed by their average gray level. The extracted region boundaries align well with the perceived boundaries at different scales. To obtain the segmentations, all parameters of the transform are selected automatically, eliminating the need to make *a priori* assumptions about either the geometric or homogeneity characteristics of the structure.

3.2 Region Matching

The matching of motion regions across frames is formulated as a graph matching problem at four different scales where scale refers to the level of detail captured by the image segmentation process. Three partitions of each image are created by slicing through the multiscale pyramid at three preselected values of σ_g. Region partitions from adjacent frames are matched from coarse to fine, with coarser scale matches guiding the finer scale matching. Each partition is represented as a region adjacency graph, within which each region is represented as a node and region adjacencies are represented as edges. Region matching at each scale consists of finding the set of graph transformation operations (edge deletion, edge and node matching, and node merging) of least cost that create an isomorphism between the current graph pair. The cost of matching a pair of regions takes into account their similarity with regard to area, average intensity, expected position as estimated from each region's motion in previous frames, and the spatial relationship of each region with its neighboring regions.

(a) Original image (b) $\sigma_g = 6$ (c) $\sigma_g = 20$ (d) $\sigma_g = 40$

Figure 2. Results of multiscale segmentation

Once the image partitions at the three different homogeneity scales have been matched, matchings are then obtained for the regions in the first frame of the frame pair that were identified by the motion segmentation module using the previous frame pair. The match in the second frame for each of these motion regions is given as the union of the set of finest scale regions that comprise the motion region. This gives a fourth matched pair of image partitions, and is considered to be the coarsest scale set of matches that is utilized in affine estimation. The details of the algorithm can be found in [12].

3.3 Affine Transformation Estimation

For each pair of matched regions, the best affine transformation between them is estimated iteratively. Let R_i^t be the ith region in frame t and its matched region be R_i^{t+1}. Also let the coordinates of the pixels within R_i^t be (x_{ij}^t, y_{ij}^t), with $j = 1 \ldots |R_i^t|$ where $|R_i^t|$ is the cardinality of R_i^t, and the pixel nearest the centroid of R_i^t be $(\bar{x}_i^t, \bar{y}_i^t)$. Each (x_{ij}^t, y_{ij}^t) is mapped by an affine transformation to the point $(\hat{x}_{ij}^t, \hat{y}_{ij}^t)$ according to

$$\begin{aligned} \begin{pmatrix} x_{ij}^t \\ y_{ij}^t \end{pmatrix} &\rightarrow R\left[\mathbf{A_k} \begin{pmatrix} x_{ij}^t - \bar{x}_i^t \\ y_{ij}^t - \bar{y}_i^t \end{pmatrix} + \vec{T}_k + \begin{pmatrix} \bar{x}_i{}^{t+1} \\ \bar{y}_i{}^{t+1} \end{pmatrix} \right] \\ &= \begin{pmatrix} \hat{x}_{ij}^t \\ \hat{y}_{ij}^t \end{pmatrix}_k \end{aligned} \tag{4}$$

where the subscript k denotes the iteration number, and $R[\cdot]$ denotes a vector operator that rounds each vector component to the nearest integer. The affine transformation comprises a $2x2$ deformation matrix, $\mathbf{A_k}$, and a translation vector, $\vec{T}_k$. By defining the indicator function,

$$\lambda_i^t(x,y) = \begin{cases} 1, (x,y) \in R_i^t \\ 0, else \end{cases} \tag{5}$$

the amount of mismatch is measured as

$$\begin{aligned} (M_i^t) = \textstyle\sum_{x,y} |I_t(x,y) - I_{t+1}(\hat{x},\hat{y})| \cdot \\ \left[\lambda_i^t(x,y) + \lambda_i^{t+1}(\hat{x},\hat{y}) - \lambda_i^t(x,y) \cdot \lambda_i^{t+1}(\hat{x},\hat{y}) \right] \end{aligned} \tag{6}$$

The affine transformation parameters that minimize M_i^t are estimated iteratively using a local descent criterion [12].

3.4 Motion Field Integration

The computed affine parameters give a motion field at each of the four scales. These motion fields are then combined into a single motion field by taking the coarsest motion field and then performing the following computation recursively at four scales. At each matched region, the image prediction error generated by the current motion field and the motion field at the next finer scale are compared. At any region where the prediction error using the finer scale motion improves by a significant amount, the current motion is replaced by the finer scale motion. The result is a set of "best matched" regions at the coarsest acceptable scales.

3.5 Motion Field Segmentation

The resulting motion field $\vec{M}_{t,t+1}$ is segmented into areas of similar motion. We use a heuristic that considers each pair of best matched regions, R_i^t and R_j^t, which share a common border, and merges them if the following relation is satisfied for all (x_{ik}^t, y_{ik}^t) and (x_{jl}^t, y_{jl}^t) that are spatially adjacent to one another:

$$\frac{||\vec{M}_{t,t+1}(x_{ik}^t, y_{ik}^t) - \vec{M}_{t,t+1}(x_{jl}^t, y_{jl}^t)||}{max(||\vec{M}_{t,t+1}(x_{ik}^t, y_{ik}^t)||, ||\vec{M}_{t,t+1}(x_{jl}^t, y_{jl}^t)||)} < m_{\sigma_g} \tag{7}$$

where m_{σ_g} is a constant less than 1 that determines the degree of motion similarity necessary for the regions to merge.

The segmented motion regions are each represented in $MS_{t,t+1}$ by a different value. Because each of the best matched regions have matches, the matches in frame $t+1$ of the regions in $MS_{t,t+1}$ are known and comprise the coarsest scale regions that are used in the affine estimation module for the next frame pair.

It should be noted that the motion segmentation does not necessarily correspond to the moving objects in the scene because the motion segmentation is done over a single motion field. Nonrigid objects, such as humans, are segmented into multiple, piecewise rigid regions. In addition, fast objects moving at rates less than one pixel per frame cannot be identified. Handling both these situations requires examining the motion field over multiple frames.

Figure 4 shows frames from an image sequence of ASL sign "any" and Figure 5 shows the results of motion segmentation. Different motion regions are displayed with different gray levels. Notice that there are several motion regions within the head and palm regions because these piecewise rigid regions have similar motion.

4 Color Segmentation

Motion segmentation generates regions that have similar motion. However, only some of these regions carry important information for gesture recognition. For example, regions of head or body movements are usually not important in recognizing hand gestures. Thus, it suffices to extract the motion of head and palm regions. We use color segmentation to find motion regions that have skin-like color. Although color feature alone is not sufficient or robust to identify human head in an image, many studies have shown that color cues, when combined with other information, help in tracking human heads and hands [7], [9], [15]. These studies use either normalized RGB or Hue-Saturation-Value (HSV) color representations to search for skin regions.

To locate human head and hands, we use CIE LUV color space in order to minimize the dependence on luminance. A look-up table is created after statistical analysis of training images that have skin-like regions. Figure 3 shows that the regions of skin color fall into a small area of LUV color space, i.e., only a few of all possible colors actually occur in human skins.

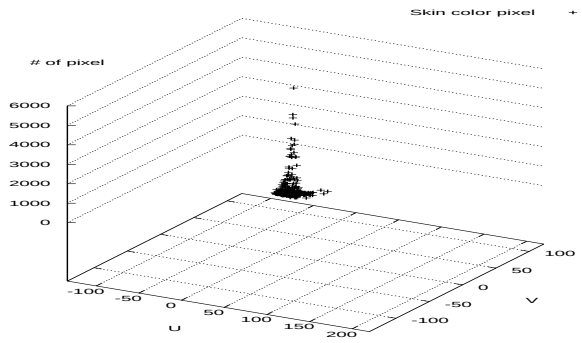

Figure 3. Characteristics of skin regions in LUV color space

A region is classified to have human skin color if most of its pixels fall into the trained LUV color space. Motion regions of skin color can be thus extracted from motion segmentation. In turn, regions of human head, hands and palms can be identified from these motion regions of skin color.

5 Geometric Analysis

Since the shape of human head and palm can be approximated by ellipses, and the human hand is a thin, rectangular region, motion regions that have skin color are merged until the shape of the merged region is approximately elliptic or rectangular.

The parameters of a rectangular fit can be obtained from the bounding box of the region easily. The orientation of an ellipse can be calculated by the moments of inertia:

$$\theta = \frac{1}{2} \cdot arctan(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}) \tag{8}$$

where $\mu_{i,j}$ denotes the second moments of normalized coordinates with respect to the centroid defined as

$$\mu_{2,0} = \int \int_{I'} (x^{'})^2 dx^{'} dy^{'} \tag{9}$$

$$\mu_{1,1} = \int \int_{I'} (x^{'} y^{'}) dx^{'} dy^{'} \tag{10}$$

$$\mu_{0,2} = \int \int_{I'} (y^{'})^2 dx^{'} dy^{'} \tag{11}$$

where $x^{'} = x - \bar{x}$, $y^{'} = y - \bar{y}$ ($\bar{x}$, $\bar{y}$ are the centroid coordinates).

The extents of the major and minor axes of the ellipse are approximated by the extents of the region along the axis directions and the degree of fit of the ellipse is determined by the number of pixels that fall into that shape specified by the computed parameters. That elliptic region which is larger than the rest is viewed as human head and the palm regions are the elliptic regions that are smaller than the head region while the hand is a rectangular area with its size in between. Figure 6 shows the results of color segmentation and geometric analysis on the motion regions.

6 Motion Trajectory

Although motion segmentation generates the affine transformations that capture motion details by matching finest regions, it is sufficient to use the coarser motion trajectories of identified palm regions for gesture recognition.

Affine transformation of palm region in each frame pair is computed based on equation (4) in Section 3.3. These affine transformations are then concatenated to construct the motion trajectory of the palm region. Figure 7 shows the motion trajectories of the palm region from the image sequence "any."

7 Experimental Results

We used a video database of more than 200 ASL signs for experiments. Each video consists of an ASL sign which

lasts about 3 to 5 seconds at 30 frames per second. Figure 4 shows some images from a video of ASL sign "any."

Figure 5 shows the results of motion segmentation. Note that the head and palm of the signer consist of several motion regions because motion segmentation is done over a single motion field as discussed previously.

Motion regions with skin color are extracted using chromatic characteristics. These regions are then merged into palm and head regions using geometric properties as shown in Figure 6.

Affine parameters are computed for each pair of segmented palm regions in successive frames. Then, the corresponding palm pixels in successive frames are connected to obtain motion trajectories for all palm pixels across the sequence. Figure 7 shows the motion trajectories of palm region in the image sequence shown in Figure 4. These trajectories match the perceived palm motion.

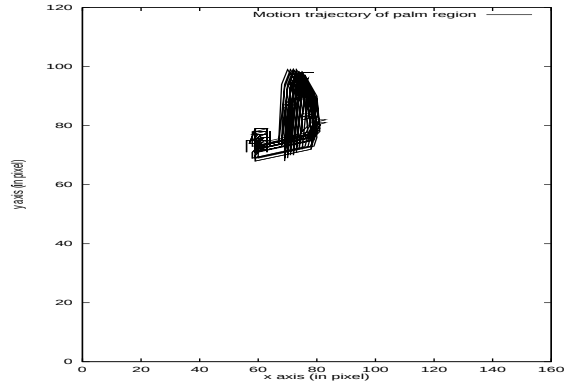

Figure 7. Motion trajectories for the palm

8 Conclusion

In this paper, we have demonstrated a method for extracting human motion trajectories based on motion segmentation, color segmentation, and geometric analysis. Experimental results show that the extracted motion trajectories match the perceived real motion. Such motion trajectories carry important spatio-temporal information that can be utilized to classify and recognize ASL gestures which is a part of our ongoing work.

Acknowledgments

The support of the Office of Naval Research under grant N00014-96-1-0502 is gratefully acknowledged. Thanks are also due to Dr. Sherman Wilcox at University of New Mexico for providing the ASL video database.

References

[1] N. Ahuja. A transform for multiscale image segmentation by integrated edge and region detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(12):1211–1235, 1996.

[2] R. Battison. *Lexical Borrowing in American Sign Language*. Linstok Press, Silver Spring, MD, 1978.

[3] Y. Cui and J. J. Weng. Hand segmentation using learning-based prediction and verification for hand sign recognition. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 88–93, 1996.

[4] T. Darrell, I. Essa, and A. Pentland. Task-specific gesture analysis in real-time using interpolated views. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 18(12):1236–1242, 1996.

[5] S. S. Fels and G. E. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Trans. Neural network*, 4:2–8, January 1993.

[6] W. T. Freeman and C. D. Weissman. Television control by hand gestures. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pages 179–183, 1995.

[7] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multimodal system for locating heads and faces. In *Proceedings of the Second IEEE International Conference on Automatic Face and Gesture Recognition*, pages 88–93, 1996.

[8] J. Lee and T. L. Kunii. Model-based analysis of hand posture. *IEEE Computer Graphics and Applications*, pages 77–86, 1995.

[9] N. Oliver and A. Pentland. Lafter: Lips and face real-time tracker. In *Proceedings of Computer Vision and Pattern Recognition*, pages 123–29, 1997.

[10] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 19(7):677–695, 1997.

[11] T. E. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proceedings of the International Symposium on Computer Vision*, 1995.

[12] M. Tabb and N. Ahuja. 2-d motion estimation by matching a multiscale set of region primitives. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 1997. submitted.

[13] M. Tabb and N. Ahuja. Multiscale image segmentation by integrated edge and region detection. *IEEE Trans. Image Processing*, 6(5):642–655, 1997.

[14] A. D. Wilson and A. F. Bobick. Recognition and interpretation of parametric gesture. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 329–336, 1998.

[15] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings of the Third Workshop on Applications of Computer Vision*, pages 142–147, 1996.

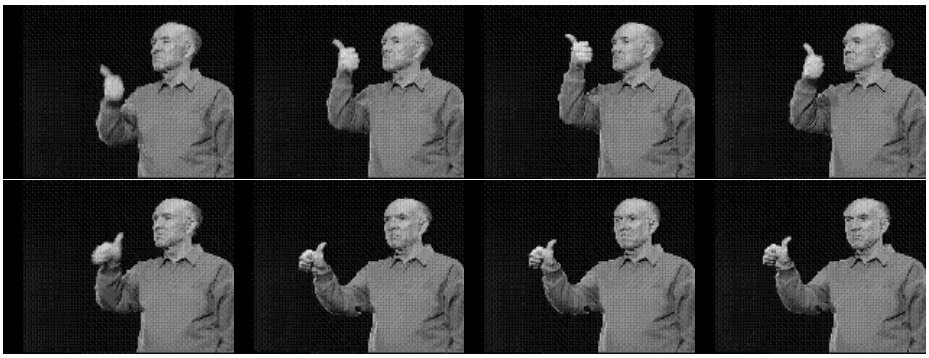

Figure 4. Image sequence of ASL sign "any" (time increases left to right and top to bottom)

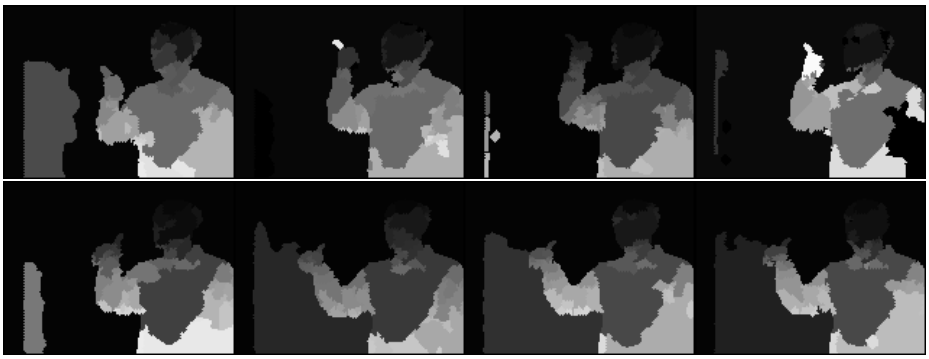

Figure 5. Motion segmentation of ASL sign "any" (time increases left to right and top to bottom)

Figure 6. Extracted human head and palm regions in the sequence of Figure 4