Novel Spatio-temporal Features for Fingertip Writing Recognition in Egocentric Viewpoint

Muhammad Zaid Hameed
Imperial College London
muhammad.hameed13@imperial.ac.uk

Guillermo Garcia-Hernando
Imperial College London
g.garcia-hernando@imperial.ac.uk

Tae-Kyun Kim
Imperial College London
tk.kim@imperial.ac.uk

Abstract

In this paper, we propose a novel feature extraction scheme for fingertip writing recognition in the air for egocentric viewpoint. The inherent challenges in the egocentric vision e.g. rapid camera motion and object's appearance and disappearance in scene may cause the fingertip to be detected in non-uniformly time separated frames. Most existing approaches do not consider this missing temporal information for feature extraction, which could be utilized to improve performance in ego-vision tasks. The novel feature extraction scheme extracts spatio-temporal features from trajectory of hand movement which are used with Hidden Markov Models for classification. The proposed feature set outperforms current trajectory based feature schemes and achieves 96.7% *recognition rate on a novel fingertip trajectory dataset.*

1 Introduction

The Egocentric vision (first person view) from wearable camera devices has constraints different from traditional vision tasks e.g. rapid camera motion, drastic changes in illumination, object's appearance and disappearance in scene which makes it more challenging and difficult for recognition tasks. At the same time, more frequent presence of hands [1, 2, 3] in the observation scene can be utilized for first person activity recognition [4, 5, 6] and hand-eye coordination [7].

Exploring new paths in terms of communicating with these wearable cameras is an interesting problem as they lack keyboard or other input devices. Writing alphabet characters using the fingertip can be one interesting way of text input. Capturing the movement described by fingertips in the air leads to the study of spatio-temporal trajectories. This trajectory representation of sequential data captures the perception of motion and orientation and thus it is extensively used as distinguishable feature representation not only for hand gesture recognition [8, 9], but also for handwriting recognition, signature verification [10] and action representation techniques.

The state of the art location, velocity and orientation features [8, 9, 11] are found to be the most promising features representing the trajectories and are widely used in hand gesture trajectory recognition problems. These trajectory feature extraction schemes are presented for third person view and consider only the trajectory data coordinates (x-y coordinates), assuming uniform time sampling. But, inherent challenges in the egocentric vision cause the fingertip to be detected in non-uniformly time separated frames. Thus, the resulting trajectory does not contain uniformly sampled trajectory data. These approaches do not consider this missing temporal information for feature extraction and a comparison of feature extraction schemes in [12] shows that trajectory based representation gives erroneous recognition results when hands are occluded during the tracking which results in a trajectory with missing information (data points). This occlusion also tends to occur in egocentric view and thus underlines the requirement of a new feature representation.

In this paper, we propose a new trajectory feature extraction scheme which incorporates temporal information during the feature extraction phase and extracted features will be used in proposed gesture recognition system. Section 2 describes the gesture recognition system along-with proposed feature extraction scheme. Performance evaluation of proposed system has been discussed in section 3. In the end this paper is concluded in section 4.

2 Gesture Recognition System

The complete gesture recognition system comprises of three stages as shown in figure 1.

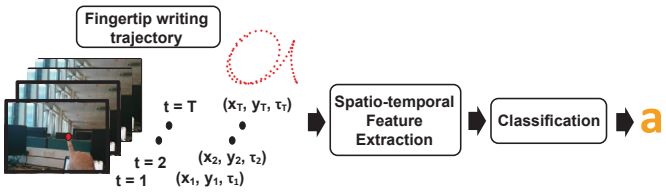

Figure 1. Proposed fingertip writing recognition system

The first stage in figure 1 consists of hand detection, fingertip detection and tracking using a depth sensor. Hand segmentation is performed with depth thresholding. For fingertip detection and tracking, contour curvature based approach [13] or distance metric approach [14] can be used which utilize the geometrical properties of hand for fingertip localization. So, spatio-temporal trajectories have been acquired by tracking fingertip over time. For comparison purpose of proposed feature set with other feature representations this part will be same and these spatio-temporal trajectories will be used for extracting features.

2.1 Proposed Spatio-temporal Feature Set

Now, we present a novel spatio-temporal trajectory feature extraction scheme that uses trajectory data coordinate information and temporal information. The hand gesture trajectory is determined by connecting the fingertip points detected in a video. Fingertip points usually have abrupt shifts in location due to shaking of hand in movement and cluttered background. In order to overcome these abrupt changes, mean of a fingertip point is computed with respect to its neighbouring trajectory points. The resulting smoothed trajectory is represented by

$$L_t = (x_t, y_t, \tau_t), \quad (t = 1, 2, ..., T) \tag{1}$$

Here, 't' represents the time frame instant and 'τ_t' represents the actual time information and 'T' completion time frame of a particular gesture.

2.1.1 Feature Extraction

The proposed feature representation includes the 2-d trajectory shape information (e.g. feature 5) similar to [8, 9, 11] and augments it by including features with temporal information efficiently in a novel way. This feature extraction process considers three consecutive points (p_t, p_{t+1}, p_{t+2}) on the spatio-temporal trajectory at a particular time instant 't' to compute spatio-temporal (ST) features and space-only features. These three consecutive points on the ST-trajectory path are used to construct a plane $\mathcal{P}$. Features calculated by surface normal vectors to the points in 3-dimensions (3-d) have been shown to efficiently capture the local geometrical shape of 3-d objects such as feature set in point feature histogram for 3-d objects [15]. The proposed ST-features (features 1-4) are calculated by using normal to the plane $\mathcal{P}$ which efficiently captures the spatio-temporal characteristics of trajectory data. The extracted feature set is represented in the following.

1. **Plane to ST-Trajectory Centroid Distance (D_{PC}):** Initially vectors '**u**' and '**v**' are calculated form point p_{t+1} to p_t and p_{t+2} respectively. Then, the normal vector '**n**' to the plane is given by

$$\mathbf{n} = \mathbf{u} \times \mathbf{v} \tag{2}$$

Now consider that $c_{st} = (M_x, M_y, M_\tau) = \frac{1}{T}\left(\sum_{t=1}^{T} x_t, \sum_{t=1}^{T} y_t, \sum_{t=1}^{T} \tau_t\right)$ denotes the spatio-temporal centroid of the trajectory. So, the distance of spatio-temporal centroid of the trajectory from the plane can be calculated as

$$D_{PC} = \mathbf{w} \cdot \frac{\mathbf{n}}{||\mathbf{n}||} \tag{3}$$

where, $\mathbf{w} = \overrightarrow{(x_{t+1}, y_{t+1}, \tau_{t+1}), (M_x, M_y, M_\tau)}$. The calculated distance D_{PC} has been shown in figure 2.

2. **Plane to Initial point of ST-trajectory Distance (D_{PS}):** Second feature to be calculated is the distance D_{PS} of plane $\mathcal{P}$ from the initial point of the trajectory 'p_1' as shown in figure 2. Here, $p_1 = I_{ST} = (I_x, I_y, I_\tau) = (x_1, y_1, \tau_1)$.

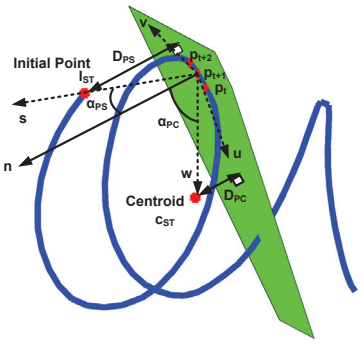

Figure 2. Computation of spatio-temporal features $\mathbf{D_{PC}}$, $\mathbf{D_{PS}}$, $\alpha_{\mathbf{PC}}$, $\alpha_{\mathbf{PS}}$

$$D_{PS} = \mathbf{s} \cdot \frac{\mathbf{n}}{||\mathbf{n}||} \tag{4}$$

here, $\mathbf{s} = \overrightarrow{(x_{t+1}, y_{t+1}, \tau_{t+1}), (I_x, I_y, I_\tau)}$.

3. **Angle of Centroid of ST-Trajectory from Plane Normal (α_{PC}):** The third feature to be calculated is the angle of centroid point 'c_{st}' from the normal vector **n** to the plane $\mathcal{P}$. This angle is shown in figure 2 and calculated using

$$\alpha_{PC} = \arccos\left(\frac{\mathbf{n} \cdot \mathbf{w}}{||\mathbf{n}||\,||\mathbf{w}||}\right) \tag{5}$$

4. **Angle of Initial Point of ST-Trajectory from Plane Normal (α_{PS}):** This feature is the angle between the normal vector '**n**' to plane $\mathcal{P}$ and the initial point of the trajectory '$p_1 = I_{ST}$' (in figure 2), computed using

$$\alpha_{PS} = \arccos\left(\frac{\mathbf{n} \cdot \mathbf{s}}{||\mathbf{n}||\,||\mathbf{s}||}\right) \tag{6}$$

5. **Spatial Tangent Angle (θ_{tan}):** The spatial tangent angle θ_{tan} is the highly discriminative 2-d trajectory shape capturing feature, calculated between points p_t and p_{t+2} in the current window of selected three points on the trajectory. The spatial tangent angle (θ_{tan}) is shown in figure 3.

$$\theta_{tan} = \arctan\left(\frac{y_{t+2} - y_t}{x_{t+2} - x_t}\right) \tag{7}$$

6. **Spatial Centroid Angle (θ_C):** The spatial centroid angle (shown in figure 3) is the angle in 2-d plane between points c_{st}, p_t and p_{t+2}, calculated using vectors '**a**' and '**b**',

$$\begin{aligned} \mathbf{a} &= \overrightarrow{(x_t, y_t), (x_{t+2}, y_{t+2})} \\ \mathbf{b} &= \overrightarrow{(x_t, y_t), (M_x, M_y)} \\ \theta_C &= \arccos\left(\frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||\,||\mathbf{b}||}\right) \end{aligned} \tag{8}$$

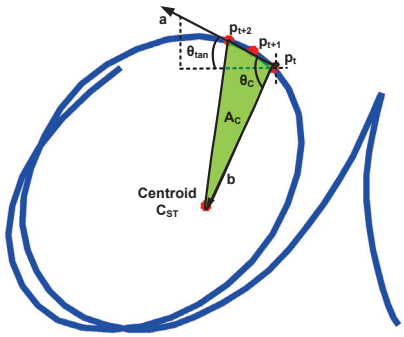

Figure 3. Space-only features $\theta_{\mathbf{tan}}$, $\theta_{\mathbf{C}}$, $\mathbf{A_C}$

7. **Centroid Triangle Area (A_C):** The last spatial feature calculated is the area of the triangle formed by points p_t, p_{t+2} and centroid c_{st} in 2-d trajectory path as shown in figure 3, which can be calculates as follows,

$$A_C = \frac{1}{2} \cdot abs \left(\begin{vmatrix} x_t & y_t & 1 \\ x_{t+2} & y_{t+2} & 1 \\ M_x & M_y & 1 \end{vmatrix} \right) \quad (9)$$

For a given character trajectory, each feature vector among features 1-7 is normalized to have values in range [0, 1] and then different weights are assigned to different features. Features 1-2 and 5-6 are experimentally proved to be more discriminative than other features and assigned a weight of 1.5 while features 3-4 and 7 are assigned a weight of 0.5 to give maximum recognition accuracy. The combined feature space representation of trajectory is given by,

$$\mathbf{F_{combined_t}} = [D_{PC_t}, D_{PS_t}, \alpha_{PC_t}, \alpha_{PS_t}, \theta_{tan_t}, \theta_{C_t}, A_{C_t}], \quad 1 \leq t \leq T-2 \quad (10)$$

2.1.2 Gesture Classification

In the third stage, The sequential data features extracted from hand movement trajectory are used for classification among different gesture classes. Hidden Markov Models (HMM) [9, 11] are generative models widely used for sequential data modelling. In this work, we have implemented a HMM system with continuous observation model for classification instead of HMM with discrete observation symbols. The reason for continuous observation model lies in the fact that critical multidimensional feature space information may be lost by vector quantization. So, HMM based classification system is implemented for performance evaluation of proposed trajectory feature set. For comparison, we have also tested our approach with a discriminative classifier, Random Forest using majority-vote rule [16].

3 Experimental Results

3.1 Dataset

The fingertip writing trajectory dataset is generated from head-mounted depth camera sensor (Creative* Interactive Gesture Camera) to record gestures in egocentric view. The dataset consists of total 260 character trajectories recorded and manually segmented from the same person, representing the 26 English alphabets (from a to z), i.e. 10 character trajectory examples for each alphabet character.

The feature vectors will be extracted from these character trajectories and will be different in size for multiple realizations of a particular gesture depending upon the variations in velocity of hand motion as well as shape and complexity of gesture. A data aligning algorithm presented in [11] is used to make feature vectors of a particular gesture of equal length.

3.2 Recognition System

For classification, a continuous HMM model is trained for every class. Output probabilities are modelled with Gaussian Mixture Models (GMM) and the training is performed using Baum-Welch (BW) algorithm [17]. The number of mixtures of Gaussian and the number of states are adjusted experimentally to 3 and between 4 and 8 respectively. 10-fold cross validation scheme has been used for performance estimation. The recognition ratio [9], for gestures of all classes can be defined as

$$R = \frac{\text{Number of correctly recognized gestures}}{\text{Total number of test gestures}} \times 100 \quad (11)$$

For recognition rate comparison, state of the art location, angle and velocity features [8] and a trajectory based writing verification feature extraction scheme [10] have been used with the same generated character trajectory dataset and the result is shown in table 1 for HMM based classification system and a discriminative Random Forest based approach. From results, it is clear that proposed feature set gives the highest recognition rate when it is used with HMM based classifier. The recognition rate of proposed feature set with Random Forest based classification is slightly lower. This is due to high temporal variation in same gesture class induced by temporal information in proposed feature set which results in lower recognition rate when used with Random Forest based approach. Also, HMM captures the sequential behaviour of the data better than Random Forest based classifier.

Table 1. Percentage recognition rate of gesture classification system

Feature Set	HMM	Random Forest
Location, angle and velocity [8]	94.6%	79.61%
Function based feature vectors [10]	85.8%	78.46%
Proposed Spatio-temporal feature set	**96.7%**	81.15%

In table 2, we present the recognition rate of individual features showing that our proposed combination of individual features leads to a significantly higher recognition accuracy. Weights of the proposed features have been adjusted experimentally so that the combined feature set gives the maximum recognition rate with classification systems.

Table 2. Percentage recognition rate of proposed individual features with HMM based system

Feature	Recognition Rate	Feature	Recognition Rate
D_{PC}	31.92%	D_{PS}	27.30%
α_{PC}	13.07%	α_{PS}	12.69%
θ_{tan}	64.23%	θ_C	41.53%
A_C	25.00%		

3.3 Gesture Recognition in Noisy Environment

In this section, performance of proposed feature set has been evaluated in noisy environment due to inherent challenges in egocentric viewpoint and two analysed scenarios have been presented in the following,

- **Case-I: Random Sample Loss of Fingertip Trajectory Data:** It is considered in the first case that fingertip is not detected multiple times in random frames during the fingertip writing process and the resulting trajectory of the data for 40% random sample loss is shown in figure 4(b).
- **Case-II: Random Sequence of Samples Loss for Fingertip Trajectory Data:** This case occurs due to occlusion in egocentric view which occurs when some object suddenly appears in the scene and occludes the fingertip in the scene for some duration randomly during gesture movement and resulting character trajectory after 12% sequential sample loss is shown in figure 4(c).

(a) Trajectory of character 'a' without any loss.

(b) Case-I, trajectory of character 'a' with 40% random sample loss.

(c) Case-II, trajectory of character 'a' with 12% sequential sample loss.

Figure 4. Character 'a' trajectories in noisy environment

The recognition result for both cases is shown in figures 5 and 6 for all trajectory data examples. The analysis of above two scenarios shows the robustness of proposed Spatio-temporal features in noisy egocentric view. It outperforms state of the art trajectory feature representation in noisy egocentric scene and maintains a high recognition rate $\approx$ 90% even for 40% random sample loss (case-I) and 15% sequence of sample loss (case-II). This is due to spatio-temporal information (features 1-4) in the proposed complete feature set (features 1-7) that not only preserves the trajectory's 2-d shape (features 5-7), but in case of high sample loss of trajectory data, it also discriminates it from other gesture classes by matching the temporal variation in the shape.

From figures 5 and 6, Recognition rate comparison of proposed 2-d space-only features (features 5-7) and complete set of proposed features (features 1-7) proves that by combining the spatio-temporal features (D_{PC}, D_{PS}, α_{PC}, α_{PS}), the complete proposed feature representation also includes the temporal information of trajectory samples obtained, which makes the feature representation resilient to high sample loss of trajectory data.

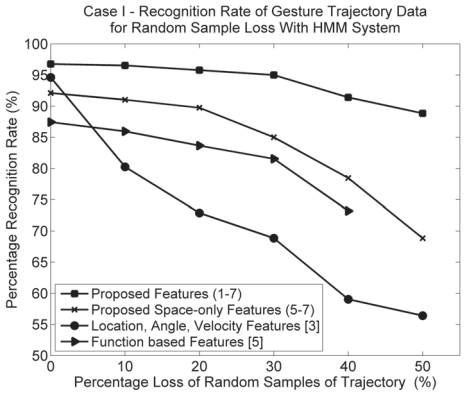

Figure 5. Character recognition rate comparison in noisy environment: case I

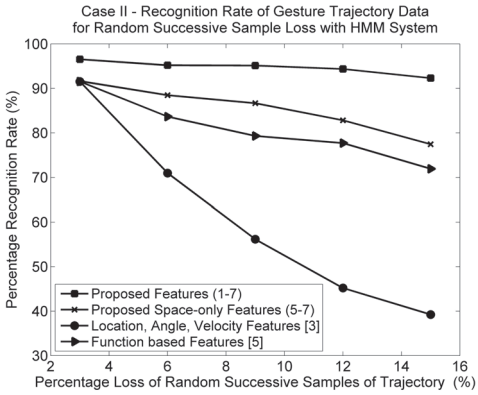

Figure 6. Character recognition rate comparison in noisy environment: case II

4 Conclusion

In this paper, a novel trajectory feature extraction scheme has been proposed which takes into account exact time information associated with 2-d trajectory data. The HMM based system implemented with the proposed feature set outperforms the recognition rate of state of the art trajectory feature sets. Finally, gesture recognition problem with sample loss of trajectory data (due to inherent challenges in egocentric view) is considered and two cases have been analysed for performance estimation of proposed feature set in noisy egocentric scene. The proposed feature set shows robustness in high noise environment and maintains a very high recognition rate as compared to state of the

art features. As a future work, we plan to extend our proposed approach to on-line egocentric handwriting detection and recognition to be able to recognize handwriting "on the fly".

References

[1] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3570–3577, IEEE, 2013.

[2] G. Rogez, M. Khademi, J. Supancic, J. Montiel, and D. Ramanan, "3d hand pose detection in egocentric RGB-D images," in *ECCV Workshop on Consumer Depth Camera for Vision (CDC4V)*, pp. 1–11, 2014.

[3] C. Li and K. M. Kitani, "Model recommendation with virtual probes for egocentric hand detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2624–2631, IEEE, 2013.

[4] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 407–414, IEEE, 2011.

[5] A. Behera, D. C. Hogg, and A. G. Cohn, "Egocentric activity monitoring and recovery," in *Computer Vision–ACCV 2012*, pp. 519–532, Springer, 2013.

[6] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2847–2854, IEEE, 2012.

[7] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Computer Vision–ECCV 2012*, pp. 314–327, Springer, 2012.

[8] H.-S. Yoon, J. Soh, Y. J. Bae, and H. Seung Yang, "Hand gesture recognition using combined features of location, angle and velocity," *Pattern Recognition*, vol. 34, no. 7, pp. 1491–1501, 2001.

[9] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, IEEE, 2008.

[10] J. Fierrez, J. Ortega-Garcia, D. Ramos, and J. Gonzalez-Rodriguez, "Hmm-based on-line signature verification: Feature extraction and signature modeling," *Pattern recognition letters*, vol. 28, no. 16, pp. 2325–2334, 2007.

[11] Z. Yang, Y. Li, W. Chen, and Y. Zheng, "Dynamic hand gesture recognition using hidden markov models," in *Computer Science & Education (ICCSE), 2012 7th International Conference on*, pp. 360–365, IEEE, 2012.

[12] S. J. Mckenna and K. Morrison, "A comparison of skin history and trajectory-based representation schemes for the recognition of user-specified gestures," *Pattern Recognition*, vol. 37, no. 5, pp. 999–1009, 2004.

[13] Z. Pan, Y. Li, M. Zhang, C. Sun, K. Guo, X. Tang, and S. Z. Zhou, "A real-time multi-cue hand tracking algorithm based on computer vision," in *Virtual Reality Conference (VR), 2010 IEEE*, pp. 219–222, Ieee, 2010.

[14] H. Liang, J. Yuan, and D. Thalmann, "3d fingertip and palm tracking in depth image sequences," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 785–788, ACM, 2012.

[15] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 3212–3217, IEEE, 2009.

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.