Expert Systems With Applications 54 (2016) 208–218

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A feature covariance matrix with serial particle filter for isolated sign language recognition

Kian Ming Lim[a,*], Alan W.C. Tan[b], Shing Chiang Tan[a]

[a] Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
[b] Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia

ARTICLE INFO

Keywords:
Sign language recognition
Feature covariance matrix
Serial particle filter

ABSTRACT

As is widely recognized, sign language recognition is a very challenging visual recognition problem. In this paper, we propose a feature covariance matrix based serial particle filter for isolated sign language recognition. At the preprocessing stage, the fusion of the median and mode filters is employed to extract the foreground and thereby enhances hand detection. We propose to serially track the hands of the signer, as opposed to tracking both hands at the same time, to reduce the misdirection of target objects. Subsequently, the region around the tracked hands is extracted to generate the feature covariance matrix as a compact representation of the tracked hand gesture, and thereby reduce the dimensionality of the features. In addition, the proposed feature covariance matrix is able to adapt to new signs due to its ability to integrate multiple correlated features in a natural way, without any retraining process. The experimental results show that the hand trajectories as obtained through the proposed serial hand tracking are closer to the ground truth. The sign gesture recognition based on the proposed methods yields a 87.33% recognition rate for the American Sign Language. The proposed hand tracking and feature extraction methodology is an important milestone in the development of expert systems designed for sign language recognition, such as automated sign language translation systems.

1. Introduction

Sign language is the structured gestures that are used by the hearing impaired community to communicate. Hearing people use verbal communication, while sign language is a form of visual communication. Very few hearing people are able to communicate in sign language. Therefore, there is a need to develop a recognition system to bridge the communication gap between the hearing and hearing impaired communities. For that reason, the sign language recognition problem has been researched since 1990s. A sign is not only performed with hand motion and hand shapes (manual components), but also through facial expressions, head motion and body postures (non-manual components). In general, sign language recognition can be categorized into isolated sign language recognition and continuous sign language recognition. Isolated sign language recognition aims to recognize a single sign gesture at a time. On the other hand, continuous sign language recognition aims to recognize the complete sign gesture sentences. In this work, we focus on isolated sign recognition solely using manual components.

Sign language recognition is widely regarded as a very challenging visual recognition task (Almeida, Guimarães, & Ramírez, 2014; Fang, Gao, & Zhao, 2004; Kong & Ranganath, 2008; Ni et al., 2013; Rautaray & Agrawal, 2015). Recognition in sign language requires visual processing on hand movement, hand shape, hand orientation, etc. Therefore, research on sign language recognition has significant impact on many expert and intelligent applications such as automated sign language translation systems, information kiosk for hearing impaired people, intelligent sign language tutoring systems and human–computer interaction systems.

To this end, we propose a serial particle filter with feature covariance matrix representation for isolated sign language recognition. First, the background model is constructed via the fusion of median and mode filters on the entire video sequence to better detect the hands. Second, based on the background model, the foreground is extracted and passed on to the proposed serial particle filter to enable the tracking of the hand trajectories. Third, the hand regions based on the trajectories are extracted and the feature covariance matrix is computed. Our main target is to create a compact representation for sign language recognition that enables the development of intelligent sign language translation systems in future work. The main contributions of this work are as listed in the following:

* Corresponding author. Tel.: +606 2523066.
E-mail addresses: kmlim@mmu.edu.my (K.M. Lim), wctan@mmu.edu.my (A.W.C. Tan), sctan@mmu.edu.my (S.C. Tan).

http://dx.doi.org/10.1016/j.eswa.2016.01.047
0957-4174/

1. A fusion of median and mode filtering to better reduce noise in the background model and thereby improves hand detection.
2. A serial particle filter enabling the tracking of the movement of both the left and right hands in the isolated sign gesture video serially, and thereby reducing misdirection of target objects.
3. A covariance matrix that fuses multiple correlated features of the tracked hands in a compact representation, and offers a dimension reduction of the features.

The paper is organized as follows: Initially, related work is presented in Section 2. Next, an overview of isolated sign language recognition is provided in Section 3. This section also describes the proposed hand detection by fusion of median and mode filtering method, serial particle filter hand tracking, feature covariance matrix for hand representation and similarity computation. Experiments and discussions are presented in Section 4. In addition, the dataset used in the experiments, the experimental setups and parameter settings, the detail performance evaluation in terms of quantitative and qualitative are explored in the same section. Finally, conclusion is drawn in Section 5.

2. Related work

Previous work on sign language recognition can be categorized based on the method of data acquisition: direct measurement approaches and vision-based approaches. Direct measurement approaches (Fang & Gao, 2002; Fang et al., 2004; Fels & Hinton, 1993; Kadous, 1996; Kim, Jang, & Bien, 1996; Kong & Ranganath, 2008; Liang & Ouhyoung, 1998; Vogler & Metaxas, 1998) collect motion data using data gloves, sensors, or motion capturing systems. These equipment are able to obtain accurate spatial information of hands, wrists, fingers, and other body parts. Even though direct measurement approaches are able to acquire the sign features directly, they are obtrusive where the signer has to wear input devices which restrict the movements. Furthermore, these input devices always require complicated setups and are very costly.

For these reasons, researchers often turn to vision-based approaches. Various techniques have been researched to extract the vision-based features for sign language recognition. Almost all began by detecting the location of hands via hand tracking approaches and many such hand trackers utilize human skin color as it is unique in comparison to the other colors. For example, Chen, Fu, and Huang (2003) proposed a fusion of skin color detection, edge detection and motion detection by a logical AND operation in their hand tracking algorithm. In the work by Rautaray and Agrawal (2011), the skin color distribution is used to segment the hand based on the L*a*b color model. Zhang and Huang (2013) combined skin color and super pixel information to extract the hand region. Although skin color is easy to distinguish, hand tracking solely based on this feature may fail due to the other exposed body parts (e.g., face or the arms) with the same skin color.

Inspired by the success of filtering methods in visual tracking tasks, researchers have begun to apply it to hand tracking. Imagawa, Lu, and Igi (1998) proposed a real time system that tracks the location of hands based on skin colors linearly using a Kalman filter. They conducted recognition based on the hands blob information. In a more recent work by Gaus and Wong (2012), Kalman filter was employed to detect hand-to-head and hand-to-hand occlusion regions. Shan, Tan, and Wei (2007) adopted a mean shift embedded particle filter as a non-linear posterior density estimator for real time hand tracking. Belgacem, Chatelain, Ben-Hamadou, and Paquet (2012) likewise embedded optical flow as a penalization method into particle filter for sign language recognition. Campr et al. (2009) used a joint particle filter to calculate a combined likelihood model of hands and head with respect to others. These literatures showed that particle filter is well-suited to the hand tracking applications, given its capability to model non-linear probability distribution, although the performance is highly dependent on a suitably chosen dynamic and observation model. On the contrary, Kalman filter is only able to work with approximate Gaussians under approximate linearity. In a recent development, Hadfield and Bowden (2014) estimated motion flow in video sequences using unregularized multiple hypotheses scene particles.

Apart from hand movement, hand shape is another important feature in sign language recognition. Due to the success in speech recognition, Starner and Pentland (1997) engaged Hidden Markov Models (HMMs) for real time American Sign Language recognition from video. Grobel and Assan (1997) extracted the location, hand shape and orientation of both hands. The signer wore color gloves to ease the calculation of size and center of gravity of the area. Position of the shoulders and the central vertical axis of the body silhouette were predicted by a rule-based classifier. Based on these information, the feature vector was defined and trained with HMMs. Tanibata, Shimada, and Shirai (2002) modeled Japanese Sign Language using hand motions and hand shapes. They reported 64 out of 65 Japanese Sign Language were recognized using HMMs. A recent work by Ni et al. (2013) modified HMMs with discriminative training and tangent vectors method of manifold to achieve better sign language recognition rates. Jangyodsuk, Conly, and Athitsos (2014) proposed to utilize Histogram of Oriented Gradient for hand shape representation and they achieved an accuracy rate of 82%. In Athitsos et al. (2008), motion energy image (MEI) was proposed as a baseline method for sign language recognition, whereas Bobick and Davis (2001) proposed motion history image (MHI) as a temporal template to recognize the human movement.

More recently, Kelly, McDonald, and Markham (2011) proposed a weakly supervised multiple instance learning density matrices algorithm for learning and recognizing signs. Other than that, some literatures employed directed graph methods such as neural network and Bayesian network for sign language recognition. Based on motion trajectories, Yang, Ahuja, and Tabb (2002) extracted and classified two-dimensional motion in video. In their work, Time Delay Neural Network was applied to the motion patterns learned from the motion trajectories. Paulraj et al. (2008) performed Discrete Cosine Transform on the sign language video sequence to extract the features. A simple neural network model was employed on 32 Malaysian Sign Language. In other works (Karami, Zanj, & Sarkaleh, 2011; Lee & Tsai, 2009; Munib, Habeeb, Takruri, & Al-Malik, 2007), the authors employed neural networks to recognize the sign language. Suk, Sin, and Lee (2010) produced a gesture model based on the skin and motion tracking information. Then, the model was passed into dynamic Bayesian network for inferencing. Later on, Bowden, Windridge, Kadir, Zisserman, and Brady (2004) proposed a two-stage classification procedure to achieve high classification rates. In the first stage, high level description of hand shape and hand motion was extracted. Next, Markov chains with independent component analysis were used in the second stage.

Recent researches focus on the color and depth information acquired using RGB-D sensors. Almeida et al. (2014) explored the phonological structure of the language and utilized seven vision-based features based on RGB-D sensor data. Zhang, Zhou, and Li (2015) used histogram of oriented displacement to obtain the hand trajectories, and multi-SVM for classification on the sign language recorded using Microsoft Kinect. Likewise, Sun, Zhang, and Xu (2015) collected an American Sign Language dataset using Microsoft Kinect sensor. In their work, a binary latent variable was first assigned to each frame in training videos for indicating its discriminative capability. Then, a latent support vector machine model was proposed to classify the signs based on both color