DEPARTEMNT OF SCIENTIFIC COMPUTING
FACULTY OF COMPUTER & INFORMATION SCIENCES
AIN SHAMS UNIVERSITY

3D CONTINUOUS REAL-TIME ARABIC SIGN LANGUAGE RECOGNITION

A Thesis Submitted to the Department of Scientific Computing, Faculty of Computer & Information sciences Ain Shams University, in the Partial Fulfillment of the requirements for PhD Degree of Computer and Information Sciences

BY

AHMED SAMIR ELONS HUSSEIN

Master. Degree in Scientific Computing Department,
Faculty of Computer & Information sciences
Ain Shams University.

SUPERVISED BY

Prof. Dr. MOHAMMED FAHMY TOLBA

Professor, Scientific Computing Department,
and former dean of Faculty of Computer & Information sciences,
Ain Shams University.

Prof. Dr. MAGDY ABOUL-ELA

professor, Sadat Academy
for Management Sciences Maadi, Cairo, Egypt

(2012)

Abstract

This thesis aims to research and develop recognition engine for Arabic Sign Language (ASL) at a level of detail necessary for recognizing signs. The translation model depends on a video tracking system which contains Arabic signs which are translated to text in real-time response.

Automated translation systems for sign languages are important in a world that is showing a continuously increasing interest in removing barriers faced by physically challenged individuals in communicating and contributing to the society and the workforce. These systems can greatly facilitate the communication between the vocal and the non-vocal communities. For the hearing-impaired, such systems can serve as the equivalent of speech-recognition systems used by speaking people to interact with machines in a more natural way. The Arabic sign language (ASL) has some characteristics which makes the translating system is very complex. First the (ASL) is a descriptive language, that the signer should describe the word to express the meaning. Secondly the right and left hands can be used interchangeably to express the same meaning. Thirdly a lot of gestures depend on the face expressions which are out of the scope of this work. ASL has about 160 postures that do not depend on the face. The first problem is how to acquire a good hand posture and motion description from a video signal. In this work, the videos are acquired using a camera, with 24 bits/pixel color resolution and 160 X 120 pixel image size 5frames/sec.

The research begins with exploring the facilities of different feature extraction methods. Many feature generation methods have been developed using Pulse-Coupled Neural Network (PCNN). Most of these methods succeeded to achieve the invariance against object translation, rotation and scaling but could not neutralize the bright background effect and non-uniform light on the quality of the generated features. To overcome the shortcomings, the research proposes a new method to enhance the features quality. The "Continuity Factor" is defined and considered as a weight factor of the current pulse in signature generation process. This factor measures the simultaneous firing strength for connected pixels. Some signs could not be classified because a single view is not enough so, 3D model was constructed using multiple views as being suggested and implemented. A novel technique to deal with pose variations in 3D object recognition is proposed. This technique uses Pulse-Coupled Neural Network (PCNN) for

image features generation from two different viewing angles. These signatures qualities are then evaluated, using a proposed fitness function. The features evaluation step is taken before any classification steps are performed. The evaluation results of the dynamic weighting factors for each camera express the features quality from the current viewing angles. The proposed technique uses the two 2D image features and their corresponding calculated weighting factors to construct optimized quality 3D features.

Then, the Gesture- Reconstruction Module is applied. This module recognizes the continuous gestures and translates them to Arabic language. Graph matching technique was applied. A decision tree-based sub-graph isomorphism algorithm was customized and implemented.

Finally, a post processing module based on Natural Language Processing (NLP) rules is proposed to detect and the correct expected errors resulting from recognition system. Popular applications for example, Optical Character Recognition (OCR), handwritten recognition, speech recognition, etc… have been researched to increase the accuracy of the recognition using NLP rules. But previous sign language recognition researchers have never explored this concept. We suggest a new hybrid semantic-oriented approach which can correct semantic level errors as well as lexical errors, and is more accurate for especially domain-specific sign language recognition error detection and correction. Through extensive experiments, it will be demonstrated the better performance of the proposed post processing approach.

Acknowledgment

I am deeply indebted to my main supervisor and the ex-dean of our faculty **Professor. Dr Mohamed Fahmy Tolba** for his great care, valuable advices, helpful guidance and providing different facilities to carry out this work.

I sincerely acknowledge **Dr. Magdy Aboul-Ela** for the supervision of this work, continuous advices, guidance and the great help in the interpretation of the results.

Thanks also are extended to all my colleagues in the scientific department. Finally I would like to express my deepest gratitude and appreciation to all my family members specially my parents for their help and support.

Ahmed Samir El-ons.

Table of Contents

List of Figures

List of Tables

Abbreviations

2D	Two Dimensional
3D	Three Dimensional
ASL	Arabic Sign Language
HMM	Hidden Marcov Model
ANN	Artificial Neural Network
PCNN	Pulse-Coupled Neural Network
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing

LSP	Lexico Semantic Pattern
ROC	Receiver Operating Characteristics
NFA	Non-deterministic Finite Automaton

Symbols

L(i)	input linking potential
F(i)	feeding potential
S	intensity of given image element
U(i)	the activation potential of neuron
θ(i)	threshold potential of neuron
αL, αF and αq	decay coefficients
β	linking coefficient
VL and VF	coefficients of the linking and threshold potential
Ysur	firing information
Yout	Neuron output
*	convolution operator
R	matrix of weight coefficients
X(i)	Output quantity based on sigmoid function.
Y(i)	Output quantity based on step-function

List of Publications

- M.F Tolba, M. S. Abdel-wahab, M. Aboul-Ela and Ahmed Samir: Image Signature Improving by PCNN for Arabic Sign Language Recognition. Canadian Journal on Artificial Intelligence, Machine Learning & Pattern Recognition Vol. 1 No. 1, March 2010.

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba." Neutralizing Lighting non-homogeneity and Background Size in PCNN Image Signature for Arabic sign language recognition" in "Neural Computing and Applications international journal" DOI : 10.1007/s00521-012-0818-4. with impact factor 0.7 in 2012.

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba." A Proposed PCNN Features Quality Optimization Technique for Cameras Weighting in Pose-Invariant 3D Arabic Sign Language Recognition" has been accepted and awaiting for publication in "Applied Soft Computing" journal with impact factor 2.6 in 2012

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba." 3D Object Recognition Using Multiple 2D Views for Arabic Sign Language" has been published in "Journal of Experimental & Theoretical Artificial Intelligence" DOI:10.1080/0952813X.2012.680073 with impact factor 0.7 in 2012.

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba " Arabic sign language continuous sentences recognition using PCNN and graph matching " Neural Computing and Applications international journal" DOI : 10.1007/s00521-012-1024-0 with impact factor 0.7 in 2012.

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba "3D Arabic sign language recognition using linear combination of multiple 2D views" Informatics and Systems (INFOS), 2012 8th International Conference on Page**(s):**6 -13 May 2012

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba " A proposed graph matching technique for Arabic sign language continuous sentences recognition" Informatics and Systems (INFOS), 2012 8th International Conference on Page(s):14 -20 May 2012

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba " Adaptive PCNN Feature Generation Model for Arabic Sign Language Recognition " accepted and awaiting for publication in IET Image Processing Journal with impact factor 6.5.

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba "Light and Background-Independent PCNN Feature Generation Model for Arabic Sign Language Recognition" submitted in International Conference on Advanced Machine Learning Technologies and Applications (AMLTA12)

- Ahmed Samir ,M. Aboul-Ela and M.F Tolba " Error Detection and Correction Approach for Arabic Sign Language Recognition " submitted in Computer Engineering & Systems (ICCES), 2012 International Conference

Chapter 1
Introduction

The most natural way human being communicates with each other is by using voice and gestures. However, the human-machine interfaces are still very primitive, thereby forcing us to adapt to the machine requirements. The use of keyboards and/or mice is non-natural communication devices for us humans. Many researchers have dedicated efforts for years to create a more natural human-machine interface. The speech recognition is being widely researched for decades. Its fundamentals and theoretical background is well studied and many commercial products have been developed. However, the gesture recognition is still much unexplored field. This lack of research is due mainly to the very high computational load needed to process efficiently video signals. Only nowadays enough powerful processors have being developed, able to process video signals in real time and/or handle with the huge amount of information necessary to store these kinds of signals. With the advancing of the computer's hardware, more sophisticated applications can be, and more natural interfaces can be built. Nowadays, Virtual Reality, Remote Operation, Robot Command, and Sign Language Translation are the applications, which take more benefit of the development of gesture recognition techniques.

We can find in the Oxford - Advanced Learner's dictionary [1] the following definition for gesture.

"Gesture - a movement of a part of the body, esp. the hand or head intended to suggest a certain meaning"[1].

From this definition we can conclude that the objective of the gesture is to allow a more complete communication, not only using voice, or when voice use becomes impossible.

1.1 Definitions

We use gestures naturally during a conversation in order to clarify or better express ideas or actions. These kinds of gestures only help the communication, being non essential to the understanding. When we think in gesture, a hand gesture is the first thing that appears in our minds; however a gesture can be performed using any part of our body. There are gestures that use mainly the head, like"yes","no","doubt", although similar ones can be done using the hands also. However, the hands still are undoubtedly the main tools used for gestural expression.

The term”gesture” is usually related to motion [2] as we can see by the dictionary’s definition. Although, from a scientific point of view, the gestures can be divided into two distinctive categories:

A-Static

In this thesis, we will call the *static gestures* as hand postures, adopting the posture definition used by Liang :

”Posture is a specific combination of hand position, orientation, and flexion observed at some time instance” [1]

Posture or static gestures are not time varying signals, so they can be completely analyzed using only one or a set of images of the hand took in a specific time. Good examples of postures are the facial information like a smile or an angry face, and the hand postures for ’OK’ or ’stop’ hand signs, which a simple picture is enough for complete understanding.

B-Dynamic

We will reserve the word ”gesture” to describe *dynamic gestures*, according Liang.

”Gesture is a sequence of postures connected by motions over a short time span.”[1]

A gesture can be thought as a sequence of postures. In a video signal the individual frames define the postures and the video sequence defines the gesture. The head ’No’ and ’Yes’, and hand ’goodbye’ or ’come here’ gestures can only be recognized taking the temporal context information, being good examples of dynamic gestures.

The gestures usually help us in the communication; however there are cases where the gestures are the only way possible to a person communicates with other. This is the case of the hear-impaired people.

The sign language is undoubtedly the most grammatically structured and complex set of human gestures. Once it is a very challenging problem, besides it has a strong social appeal, recently, many researchers have dedicated efforts to design automatic translators from sign language to text or speech [3] and vice-versa [4]. However, reliable, fast, and vocabulary complete system are not reached yet with the current technology.

Sign Languages are genuine languages, with their own grammar, and can be very different from spoken language of the country. Moreover, a Japanese signer would have so much trouble to understand a Brazilian signer as a Japanese to understand spoken Portuguese [4], once the words and grammar are completely different, see Fig 1.1.

The American Sign Language (ASL) has more than 6000 gestures and uses 26 hand postures to represent the American alphabet [5]. The Japanese Sign Language is composed by more than 8000 gestures and contains a finger spelling composed by 76 hand postures and gestures [6]. The finger spelling in sign languages is used mainly to describe names and places when a correspondent sign is not available. Not only the hands take an important place in SL recognition, head movements and facial information are very important to the understanding the sign languages as we can. Besides the facial information, the emotional expression can be added to sign by varying sign parameters such as [7]:

- speed of gesture;
- size of gesture space;
- number of repetitions or duration;
- tension of gesturing; and
- Hold-time of a posture while signing.

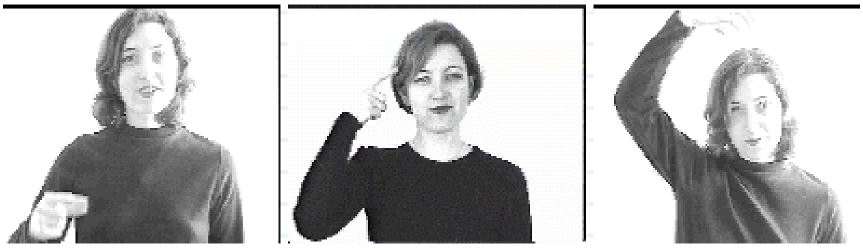

Fig. 1.1 signs of government from three different sign languages [7].

1.2 Problem Statement

Any 3-D object (hands in our case) may be represented as one or more images taken from different viewpoints. In most object recognition scenarios the object of interest is at a viewing distance that gives a clear view of the object as a whole with sufficient detail visible to render it distinctive. In such a scenario, the depth variation across the object of interest is usually sufficiently small in comparison to its distance from the camera that the perspective projection may be well-approximated by an affine projection. In a view based object recognition approach, or in other words, the problem of recognizing an object from a single 2-D image may then be formulated as follows:

> *Suppose we are given a template function (F0), a target scene image function (I) and Transformation (T) that transforms the template into (F). The goal of object recognition is to minimize*
>
> $$P = g(I,F)$$
>
> *With respect to the transformation (T). g(...) is a matching metric giving rise either to a dissimilarity or similarity score. If the minimum of (P) is smaller than or equal to some threshold (Th), then we can say we have a match.*
> *The transformation (T) can be a combination of:*
>
> *Translation-Rotation-Scaling*

The main difficulty that arises in the previous formulation is that it does not consider both lighting conditions and background size. The problem statement becomes to find the best matching within threshold value in any transformation. These transformations include lighting non-homogeneity and relatively large background size in the image scene.

Once the problem has been resolved for a single 2-D view, in some situations a single 2-D view is not sufficient to distinguish between 2 different postures. The next step is to make use of view-based approach. This involves using more than one representative view of the object at the same time. In this approach, 3-dimensional objects are represented by methods based on a combination of 2-D images.