A Kinect Based Sign Language Recognition System Using Spatio-temporal Features

Abbas Memiş, Songül Albayrak
Department of Computer Engineering, Yıldız Technical University, İstanbul, Turkey
abbasmemis@gmail.com, songul@ce.yildiz.edu.tr

ABSTRACT

This paper presents a sign language recognition system that uses spatio-temporal features on RGB video images and depth maps for dynamic gestures of Turkish Sign Language. Proposed system uses motion differences & accumulation approach for temporal gesture analysis. Motion accumulation method, which is an effective method for temporal domain analysis of gestures, produces an accumulated motion image by combining differences of successive video frames. Then, 2D Discrete Cosine Transform (DCT) is applied to accumulated motion images and temporal domain features transformed into spatial domain. These processes are performed on both RGB images and depth maps separately. DCT coefficients that represent sign gestures are picked up via zigzag scanning and feature vectors are generated. In order to recognize sign gestures, K-Nearest Neighbor classifier with Manhattan distance is performed. Performance of the proposed sign language recognition system is evaluated on a sign database that contains 1002 isolated dynamic signs belongs to 111 words of Turkish Sign Language (TSL) in three different categories. Proposed sign language recognition system has promising success rates.

Keywords: Turkish Sign Language, sign language recognition, dynamic sign gestures, spatio-temporal features, Kinect, depth maps.

1. INTRODUCTION

Human-computer interaction (HCI) based systems have been used especially in operating systems management, smart home controlling systems, human action detection and recognition systems widely in recent years. Vision based action or gesture recognition researches have become more challenging as a sub research domain of HCI along with the new real-world problems and requirements.

Sign languages are visual languages which consist of movements, locations, shapes of hands and arms; facial expressions and also enable deaf community to communicate with each other and others. Not only sign languages are native languages for deaf community, but also have national characteristics as native spoken languages. Even, some nations speak same languages have different sign languages [1]. Sign language recognition researches and studies have been done with the assistance of computer technology for a few decades. There are two main approaches, which are sensor glove and vision based approaches, for sign language recognition in terms of gesture data acquisition. Although sensor glove or electronical device based systems provide more robust and reliable data, are not user-friendly and practical like vision based systems. First vision based sign language recognition systems have been started to present in first period of 90s. Researches on these systems have been focused on different problems that are morphological and general structures such as fingerspelling, static or dynamic word representing, phrases of sign languages. In addition, signer dependency has become more important in recent studies.

Sign language words can be categorized into two sub-sections as hand postures and hand gestures. Hand postures are known as static gestures because hand postures can be represented by an image. Hand postures are usually used to describe numbers and letters of alphabets but hand gestures are commonly used to describe words or sentences. Most of the sign language words consist of dynamic hand gestures which are continuous sequences that appear with the shape and location variation of hands in time domain. Also hand gestures have more number of words than the hand postures. These cases make hand gesture recognition a more challenging problem. There are different approaches that use sensor gloves or cameras in literature for sign language recognition problem. A sensor glove based system was proposed by Öz and Leu [2], uses artificial neural network approach in order to recognize American Sign Language (ASL) words over a dataset has 50 words by success rate about %90. However, in another sensor glove based system [3], Also Gao, Fang, Zhao and Chen used self organizing maps and hidden Markov model approaches for Chinese Sign Language recognition

and observed a success rate nearly %87 by evaluating system over large dataset has 5113 signs. In vision based systems, HMMs are well known and widely used feature extraction methods because of the variation of hand location and shape in time space. Haberdar and Albayrak [4] presented a HMM based Turkish Sign Language recognition system uses local features and observed a % 95.4 recognition rate over 50 sign words. Shanableh [5] extracted spatio-temporal features by means of motion prediction error and spatial transformation for Arabic Sign Language recognition. This system was evaluated on a dataset has 23 different signs, has successful rates between %97 and %100 for different classifier models. Starner [6], proposed the first vision based sign language recognition system, recognized 494 phrases belongs to American Sign Language with %75 recognition rate on Kinect data by using HMMs.

Proposed system is presented in this paper, describes a Turkish Sign Language recognition system that uses RGB video sequences and depth maps data captured by a Kinect sensor. Spatio-temporal features are used in order to describe and recognize dynamic TSL signs.

2. TURKISH SIGN LANGUAGE DATASET

Although there are related studies on Turkish Sign Language recognition [4,7] there are not common and comprehensive datasets in TSL. In addition, some research groups in universities and associations like Turkish National Language Association have been worked on sign language dictionaries describes the characteristics and representations of signs. In the scope of proposed system, a new large dataset is collected which contains words from different categories. Related dataset contains 111 sign of words and short sentences in three categories which are about the terms of daily life, time and jobs. Some selected words of these categories are presented in table 1. Each sign in dataset is carried out with 3 repetitions of 3 different deaf people. Related dataset contains separated words and short sentences with different number of frames. A sample representation of a sign in dataset is shown in figure 1.

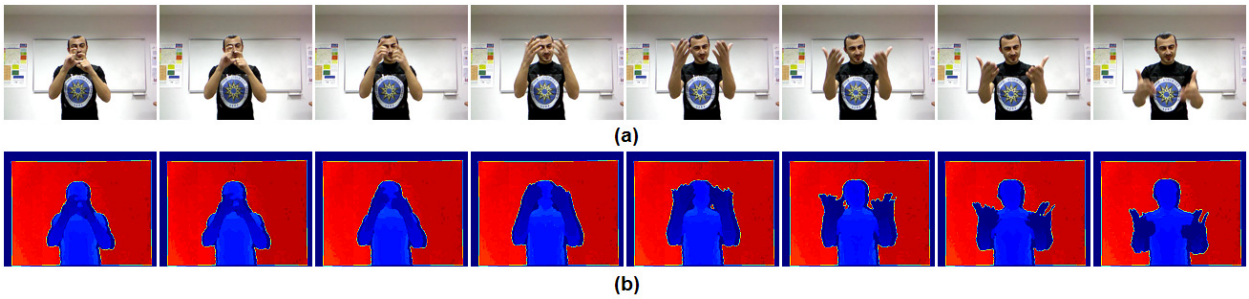

Figure 1. RGB video (a) and depth map (b) sample frames which represent “good morning” word in related sign language dataset.

Table 1. Selected word and sentence samples from TSL dataset.

Category-1 (22 words)		**Category-2 (58 words)**		**Category-3 (31 words)**	
TURKISH	**ENGLISH**	**TURKISH**	**ENGLISH**	**TURKISH**	**ENGLISH**
günaydın	good morning	bazen	sometimes	mühendis	engineer
merhaba	hello	cuma	friday	elektrikçi	electrician
teşekkürler	thank you	ocak	january	öğretmen	teacher
iyiyim	i’m fine	yaz	summer	öğrenci	student

3. PROPOSED SIGN LANGUAGE RECOGNITION SYSTEM

3.1 Video Image Preprocessing

Preprocessing steps in proposed system consists of frame resizing and color space transformation. RGB video sequences and depth maps whose original sizes are 640x480 pixel, are resized by using 1/2 and 1/4 scaling rates respectively, and 320x240 and 160x120 frame sizes are obtained. This process is employed in order to observe the effects of different frame sizes of video data on system recognition performance. Color space of RGB video sequences transformed into grayscale as the following step. Preprocessing steps are followed by the feature extraction processes.

3.2 Feature Extraction

Gestures of sign languages consist of dynamic hand gestures which are continuous sequences that appear with the shape and location variation of hands in time domain. Therefore, successive frames include valuable information that can describe gesture in time domain. Proposed system uses two-step feature extraction approach, which is based on spatio-temporal analysis, and this approach is detailed in the following sections. Temporal motion differences of successive frames are utilized in order to extract temporal features of dynamic signs in dataset. Motion differences of successive frames of the RGB video sequences and depth maps are accumulated in two accumulated motion images that represent the all motion in time space for each sign video sample. Accumulated motion images contain noises and have separated motion segments. So median filter kernel is applied to accumulated images in order to decrease the noise effects and represent motion segments as a whole. Then, 2D Discrete Cosine Transform (DCT) is applied on filtered accumulated images which contain temporal information to obtain spatial features. 2D DCT coefficients that contain higher energy of transformed images are picked up via zigzag scanning. Then, feature vectors are generated using DCT coefficients of accumulated motion images that are obtained from gray-images and depth maps.

3.2.1 Temporal Feature Extraction

Motion prediction approach in successive video frames is one of the sub-processes in video compressions. In proposed system, this approach is employed in order to represent and extract temporal features. This operation referred as in equation (1) is carried out by calculating the intensity differences of all successive frames.

$$I_{Di} = I_{i+1} - I_i; \qquad (i = 1, 2, ..., n-1) \tag{1}$$

In equation (1), I_i and I_{Di} denote the ith video frame and difference image of two successive video frames respectively. Each difference image is transformed into binary form by binary thresholding operation and then a motion accumulation image which represents the temporal features is obtained by normalizing and summing the binary images as expressed in equation (2). In equation (2), I_{Ti} denotes thresholding operation applied difference image, T denotes threshold value, w_i is the frame weight and I_A corresponds to the accumulated motion image. Proposed system uses binary thresholding operation, which eliminates the small motions differences of successive frames, in temporal feature extraction. According to the selection of the threshold value, accumulated motion images can have different motion representation. So feature vectors and also recognition performance can be affected by threshold selection. Since gray-images and depth maps have different characteristics, threshold values are determined different for each of these input data. So, threshold values are determined as automatically for each of two successive frames in a sign video sample by calculating the mean of absolute intensity differences which are non-zero.

$$I_{Ti} = \begin{cases} 1 & if\ |I_{Di}| \geq T \\ 0 & else \end{cases} \qquad I_A = \sum_{i=1}^{n-1} I_{Ti} \cdot w_i \tag{2}$$

Sign video samples in dataset do not have same size in time domain. So they should be normalized while obtaining the I_A accumulated motion images considering their sizes. In proposed system each frame difference is weighted by a fixed coefficient $w_i = 1/n$. Median filter is applied to accumulated images since motion images contain some pixels which can be denoted as noise effect and motion segments have fragmented structure. 2D accumulated motion images, which contain temporal features of signs in time domain, will be used in spatial feature extraction process. Some instances of accumulated motion images of gray-images and depth maps, which belong to "good evening" in related dataset, are shown in figure 2.

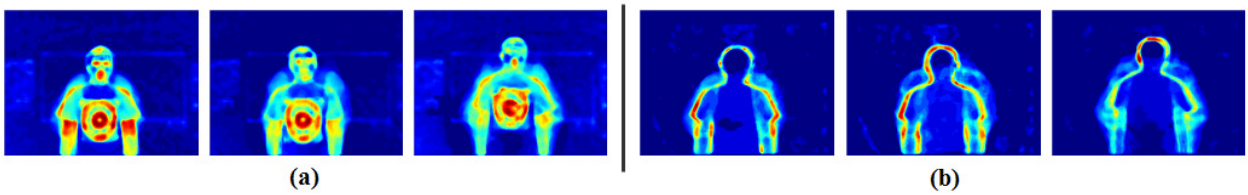

Figure 2. Accumulated motion images obtained by differences of (a) gray-images and (b) depth maps belong to to the "good evening" sign which is performed by 3 different signers (Images are normalized to 0-255 intensity range).

In accumulated motion images in figure 2, warm colors (tones of red) point out the regions that have more motion activities, and cool colors (tones of blue) point out the regions that have less motion activities.

3.2.2 Spatial Feature Extraction

After the operations about temporal feature extraction, 2D DCT is applied to accumulated motion images of gray-image and depth map in order to obtain spatial features. Temporal features of dynamic signs are transformed into spatial domain by 2D DCT. DCT provides a coefficient matrix, which has coefficients that have higher energy on the top-left corner, as an output. Also, DCT enables images to represent with lower size data and this property makes it a useful and preferred method in data compression approaches. In proposed system, 2D DCT is performed as a sub-operation to obtain coefficients that will be used to represent feature vectors. DCT coefficients in gray-image and depth map transform images picked up via zigzag scanning that starts from the top-left corner of coefficient matrix, and merged in definite rates to generate actual feature vectors.

4. EXPERIMENTAL RESULTS

A new Turkish Sign Language dataset that contains totally 111 words and sentences is collected to evaluate the performance of proposed sign language recognition system. Words or sentences in related dataset collected in three categories which are defined as terms of daily-life, time and jobs and these words selected as considering the most known words in daily-life conversations. Also each sign in dataset is performed by a deaf signer is professional and experienced at performing signs. Signs videos recorded as RGB-video sequences with their corresponding depth maps by a Kinect sensor. Each word in dataset has 9 records which are performed 3 times by 3 signers. While some samples in dataset are not used in performance evaluation since they are affected noise heavily, some 4th existing records of some signers added performance evaluation and totally 1002 samples have been evaluated.

In feature extraction step, 2D Discrete Cosine Transform is performed in order to extract the features of signs. Feature vector of each sample is constructed by composing the DCT coefficients, which are selected via zigzag scanning, of gray-image and depth map in (1:1) equal ratios. In performance evaluation step of proposed system, experiments have been done on different feature vector lengths and different video frame sizes in order to observe the effects of the number of DCT coefficients in feature vectors and frame sizes. Also it is aimed to observe and select the optimum feature vector length and video frame size. K-Nearest Neighbor (K-NN) classifier which is known as a simple classifier is performed to recognize sign samples. In K-NN classifier, which uses Manhattan distance, K parameter is determined as 1. Train and test samples in related dataset are determined by 3-fold cross validation operation. Correct recognition rates (CRRs) are considered in order to analyze the performance of proposed system. CRRs are measured as getting ratio of number of correctly classified samples to the number of all samples in dataset. In figure 3 CRRs for different coefficient ratios in 160x120, 320x240 and 640x480 frame sizes are presented. Figure 3 presents the CRRs of gray-image + depth features of all 111 words in TSL dataset.

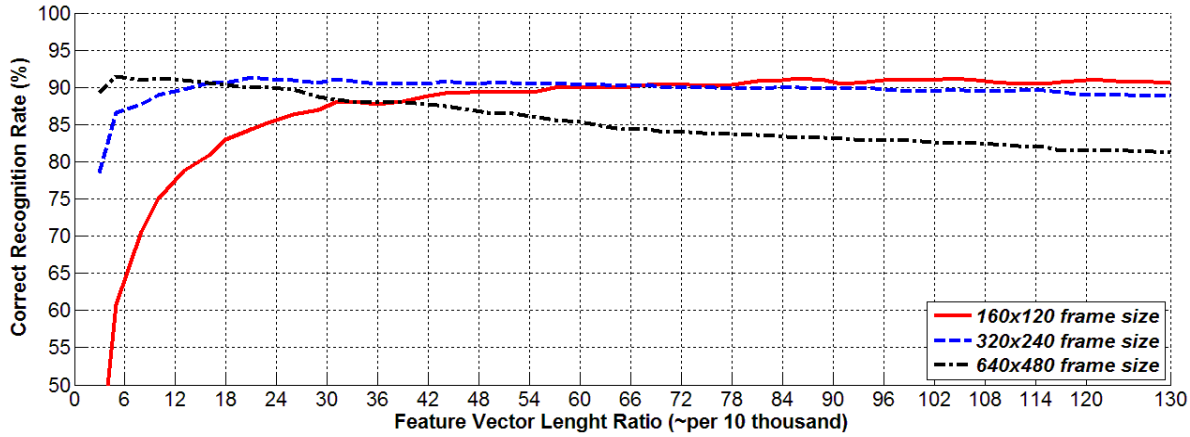

Figure 3. CRRs of proposed system for different DCT coefficient ratios in 160x120, 320x240 and 640x480 frame sizes.

In Table 2, most successful recognition performances of gray-image, depth and gray-image + depth features are presented. Table 2 shows that performances of gray-image and depth features are quite closer to each other and combination of these features provides more successful recognition rates. DCT ratio values in table 2 denote the ratio of number of DCT coefficients that are in feature vectors to the number of all DCT coefficients.

Table 2. Performances (CRRs) of gray-image, depth and gray-image + depth features on TSL dataset in different frame sizes.

	Correct Recognition Rates (%)					
	160x120		320x240		640x480	
	DCT R.	CRR	DCT R.	CRR	DCT R.	CRR
Gray Image	130	89.52	39	89.52	10	90.12
Depth	99	88.42	31	90.52	8	89.72
Gray Image + Depth	86	91.22	21	91.32	5	91.52

5. CONCLUSION

In proposed sign language recognition system on Turkish Sign Language, isolated sign samples are successfully recognized by accumulated motion image approach based on the differences of successive video frames. This simple approach enables to obtain feature vectors by transforming temporal features into spatial domain via Discrete Cosine Transform. In performance evaluation K-NN classifier that uses Manhattan distance is performed. Proposed system does not use object detection (like hand, face etc.), segmentation and tracking operations. Although this approach has advantage when these operations have been considered, environments have dynamic or moving backgrounds may affect the system efficiency and it is possible to consider segmentation and tracking approaches to overcome these kinds of problems. Also, small motion differences in successive frames are eliminated by thresholding operation on motion difference images. In thresholding operation, threshold coefficients are determined automatically by mean thresholding instead of using empirically determined threshold coefficients. Sign samples, which correspond to 111 words and sentences in related dataset which is evaluated in proposed system, were performed by deaf signers. Signs samples were captured by a Kinect sensor which has a RGB camera and two IR based sensors which enable to capture depth data in 640x480 video frame sizes. Feature vectors are constructed from gray-images that converted from RGB data and depth maps. Not only performance of feature vectors of gray-image and depth map data in merged form is evaluated, but also performances of these feature vectors are evaluated separately. CRRs obtained in performance evaluation show that depth maps contains significant information for sign or gesture recognition. Also it can be said that it is efficient to work with the frames that have lower sizes (160x120) to decrease workload as a result of experiments on sign samples.

Consequently, isolated sign samples belong to Turkish Sign Language are recognized by promising success rates. Proposed system has a maximum success rate %91.52 on overall 111 words of three categories.

ACKNOWLEDGEMENT

This research has been supported by Yıldız Technical University Scientific Research Projects Coordination Department. Project Number: 2012-04-01-YL03.

REFERENCES

[1] C. Lucas. The Sociolinguistics of Sign Languages, Cambridge University Press, Cambridge (2001).

[2] C. Öz and M.C. Leu, "American sign language word recognition with a sensory glove using artificial neural networks," Engineering Applications of Artificial Intelligence **24**, 1204-1213, (2011).

[3] W. Gao, G. Fang, D. Zhao and Y. Chen, "A chinese sign language recognition system based on SOFM/SRN/HMM," Pattern Recognition **37**, 2389-2402, (2004).

[4] H. Haberdar and S. Albayrak, "Real time isolated turkish sign language recognition from video using hidden markov models with global features," in Computer and Information Sciences – ISCIS, ser. Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg, 677-688, (2005).

[5] T. Shanableh, K. Assaleh and M. Al-Rousan, "Spatio-temporal feature extraction techniques for isolated gesture recognition in arabic sign language," IEEE Trans. on Systems, Man and Cybernatics **37**, 641-650, (2007).

[6] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton and P. Presti, "American sign language recognition with the kinect," International Conf. on Multimodal Interfaces, 279-286, (2011).

[7] O. Altun and S. Albayrak, "Turkish fingerspelling recognition system using generalized hough transform, interest regions, and local descriptors," Pattern Recognition Letters **32**, 1626-1632, (2011).