2013 ICT-PAMM Workshop on

Mobility Assistance and Service Robotics

Organized by

ICT-PAMM (France-Asia ICT Project 2011-2013)

Sponsored

Kumamoto University

Corporate sponsored by

SANWA HI-TECH Co., Ltd.

November 8-10, 2013

ASO Farm Land, Kumamoto, Japan

Proceedings of the 2013 ICT-PAMM Workshop on Mobility Assistance and Service Robotics

Message from the General Chairs of Program Committee

It is our great pleasure to welcome all of you to the Workshop on Mobility Assistance and Service Robotics (2013 ICT-PAMM), which will be hold in the period of November 8-10 at Aso Farm Land Resort Hotel, Kumamoto, Japan. This workshop is sponsored by France-Asia ICT Project ICT-PAMM and Kumamoto University.

The purpose of this workshop is to discuss topics related to the challenging problems of service robotics, mobility assistance and driving assistance in open and dynamic environments. The integral assistance systems are robotic modules and technological aids in general for personal assistance, such as robots, mobile bases, electric wheelchairs, soft robot manipulator arm. They can support disabled and elderly people with special needs in their living environment. This workshop will focus on the assistance of human in terms of its mobility and its social interaction. This workshop offers a track of quality R&D updates from key experts and provides an opportunity in bringing in the new techniques and horizons that will contribute to Mobility Assistance and Service Robotics in the next few years.

Organizing a workshop is a challenging job. We would like to express our gratitude to the members of 2013 ICT-PAMM Organizing and Program Committees who have delicately contribute their spirit and time to make such wonderful job. Finally we express our hearty congratulations to all participants, and we hope that this workshop will be an informative and memorable experience for all of you.

Prof. Nobuki Murayama
General Chair
Kumamoto University, Japan

Dr. Christian Laugier
General Chair
INRIA-eMotion, France

Date

November 8-10, 2013

Venue

ASO Farm Land

5579-3 Kawayo, Aso-gun Minamiasomura, Kumamoto, 869-1404 Japan

http://www.asofarmland.co.jp/index.php

Kumamoto University Kurokami South Campus

2-39-1 Kurokami, Chuo-ku, Kumamoto, 860-8555 Japan

http://ewww.kumamoto-u.ac.jp/

Organized by

ICT-PAMM (France-Asia ICT Project 2011-2013)

Sponsored by

Kumamoto University

Corporate Sponsored by

SANWA HI-TECH Co., Ltd.

Program Committee

General Chairs

Prof. Nobuki Murayama (Kumamoto Univ., Japan)

Dr. Christian Laugier (Inria, France)

Co-Chairs

Prof. Zhencheng Hu (Kumamoto Univ., Japan)

Dr. Anne Spalanzani (INRIA-eMotion, France)

Prof. Philippe Martinet (IRCCYN-CNRS, France)

Prof. Eric Castelli (MICA Center, Vietnam)

Dr. Fawzi Nashashibi (INRIA-IMARA, France)

Prof. Sukhan Lee (SKKU, Korea)

Prof. Ren Luo (Taiwan University, Taiwan)

Local Organizing Committee

Chair

Prof. Zhencheng Hu

Co-Chairs

Prof. Nobutomo Matsunaga

Prof. Tomohiko Igasaki

Prof. Toshiro Matsuda

Prof. Kohichi Ogata

Prof. Akio Tsuneda

Prof. Hiroshi Okajima

Prof. Masayuki Tanabe

Table of Contents

2013 ICT-PAMM Workshop on Mobility Assistance and Service Robotics

2nd Best Paper Award

is hereby granted to

T-H. Tran, V-T. Nguyen, V-N. Nguyen and Q. Midy

for outstanding research activity and presentation of

Vision based dynamic hand gesture recognition

Awarded: November 9, 2013

Prof. Nobuki Murayama
General Chair of ICT-PAMM 2013
Kumamoto University, Japan

Dr. Christian Laugier
General Chair of ICT-PAMM 2013
INRIA-eMotion, France

ISRC INRIA CNRS Kumamoto University 国立大学法人 熊本大学 MICA

Vision based dynamic hand gesture recognition

Thanh-Hai Tran, Van-Toi Nguyen, Van-Ngoc Nguyen
Computer Vision Department
International Research Institute MICA
HUST- CNRS/UMI-2954- INP Grenoble
{thanh-hai.tran, van-toi.nguyen, van-ngoc.nguyen}@mica.edu.vn

Quentin Midy
Electronic Department
ENSEIRB-MATMECA
Bordeaux, France
midy.quentin@live.fr

***Abstract*—This paper presents a comparative study on methods for dynamic hand gestures recognition. We propose 3 methods for studying: PCA-KNN, GIST-KNN, and KDES-A. The dataset we use for evaluating is a dataset of 20 Italian hand gestures provided by the big CHALEARN contest (http://gesture.chalearn.org/homewebsourcereferrals). This is a multimodal data that contains video, sound, skeletons, etc. In our work, we make our analysis on video based hand gesture recognition.**

Keywords—dynamic hand gesture recognition, PCA, KNN, Gist kernel descriptor, comparative study

I. INTRODUCTION

CHALEARN is a contest on gesture and sign language recognition from video data organized by Microsoft [1]. Last year, the contest focuses on hand gesture recognition based on only *one shot learning,* knowing that traditional recognition methods require a lot of training data (e.g. SVM, Boosting) (contest in 2011/2012). This year, the contest focuses on hand gesture using multimodal information coming from RGB-D and also audio sensors.

This paper presents our works on the current contest of hand gesture recognition from RGB-D sensor. To be able to rise up the best method for application in reality, we do a comparative study on a common and challenge data set provided by CHALEARN contest in 2013.

Our comparative study is carried out with three methods for dynamic hand gesture recognition: i) PCA – KNN; ii) GIST – KNN; iii) Kernel Descriptor. All methods are based on the pre-processing the video shot that is the computation of the Motion History Image (MHI).

II. PROPOSED METHOD

A. General framework for dynamic hand gesture recognition

As we will compare several methods on a same problem, we need to build a unified framework so that all methods will be integrated inside, taking the same input. The framework consists of two phases: learning and recognition. We can see the principal modules in both modules:

1. **Compute MHI:** As a dynamic hand gesture is a sequence of consecutive frames, we propose to represent each video shot containing one dynamic hand gesture by a Motion History Image (MHI) computed from this frame set.
2. **Feature extraction:** So feature extraction will be computed on the HMI. We can extract many types of state of the art features for example PCA components, Gradient, Texture, Color distribution or GIST features.
3. **Model learning:** In function of extracted features, a compatible recognition model will be chosen. We propose to use KNN (K-Nearest Neighbor) and Kernel descriptor. KNN is a traditional method while Kernel descriptor based method is a new trend in machine learning (Fig. 1).
4. **Recognition:** Finally to evaluate the methods, we test all examples in the testing data using learnt models previously (Fig. 2).

In the following, we will present in detail each module in the overall system.

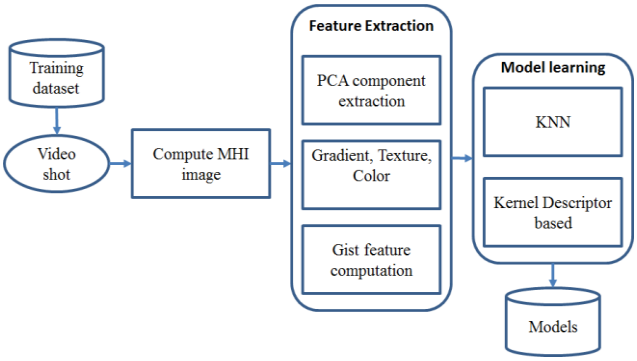

Fig. 1: Learning phase

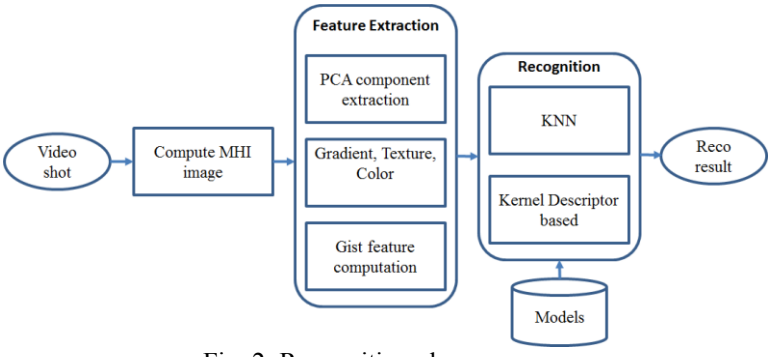

Fig. 2: Recognition phase

B. Computation of Motion History Image

As we work with dynamic hand gestures, we need to build hand gesture description based on motion information. As reported in [1], there are many features for motion representation among which Spatio-temporal features is an example (STIP) [2]. STIP has shown very success for many recognition tasks however its computation is very expensive. In addition, STIP is based on BOW model that discards geometric relationship of the features then hence is less discriminative. Hidden Markov Model (HMM) is a generative model that can deal with dynamic hand gesture representation. However, the computation time is significant.

We then follow the approach proposed in [1] in which the sequence of consecutive frames of one hand gesture is represented by only one image: Motion History Image [3]. All frames in a video sequence are projected onto one image across the temporal axis. MHI captures the temporal information of the motion in a sequence (Fig. 3).

The MHI is computed as follows: Given $I_t = \{I_1, I_2, \ldots I_{nFrames}\}$ is a grayscale image sequence. Let $B_t = \{B_1, B_2, \ldots, B_{nFrames-1}\}$ be a binary image sequence indicating regions of motion which can be obtained from image differencing and thresholding.

$$B_t = \begin{cases} 1 & \text{if } (I_{t+1} - I_t) > Threshold, \\ 0 & \text{otherwise.} \end{cases}$$

Where threshold is defined as:

$$Threshold = \sqrt{\sum_{t}^{nFrames} \sigma_t / (h \times w \times nFrames)}$$

Where σ_t is the second moment of a single frame I_t, h, w are the height and width of the video sequence. The motion history image MHI $H(t; \tau)$ is used to represent how the motion image is moving and defined as:

$$H(t;\tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ max(0, H((t-1);\tau) - 1) & \text{otherwise.} \end{cases}$$

With τ is set to be sequence duration.

Fig. 3: MHIs of two video shots in CHALEARN dataset

Finally, each video shot will be represented by an MHI which has the resolution equaling to the sequence resolution.

C. Recognition based on PCA – KNN

Given an MHI having the size of WxH elements (W, H are the width and the height of the MHI), we think to a traditional method to reduce the size of feature vector. We have used the training dataset (described in the experimental section) to build the PCA (Principal Component Analysis) space, then represent each MHI in this space by vector of 64 dimensions (this number is selected by experience). For classification of hand gesture, we use KNN to find the best match given a hand gesture.

D. Recognition based on GIST – KNN

Results of variety of state of the art scene recognition algorithms [4] shown that Gist features[1] [5] obtains an acceptable result of scene and object classification (appr. 73 – 80 %). Therefore, in this study, we propose to use Gist features to characterize MHI.

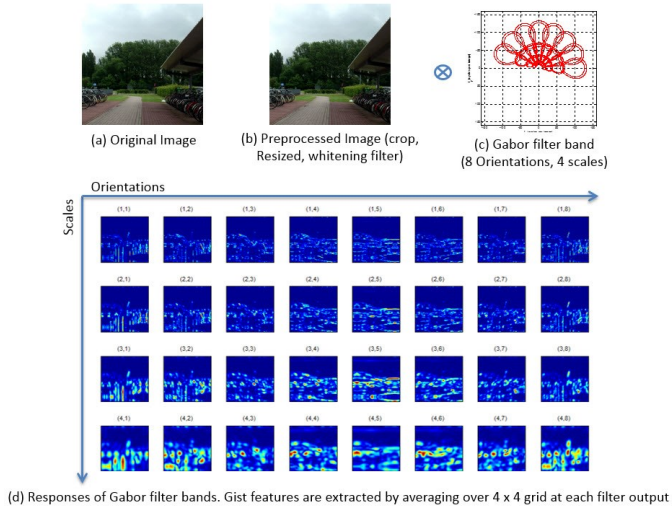

Fig. 4. *Gist feature extraction from input image*

To capture remarkable/considering of a scene, Oliva *et al.* in [5] have evaluated seven characteristics of an outdoor scene such as naturalness, openness, roughness, expansion, ruggedness, so on. They suggested that these characteristics may be reliably estimated using spectral and coarsely localized information. Steps to extract Gist features are explained in Fig. 4.

Firstly, an original image is converted and normalized to gray scale image I(x, y) (Fig. 4 (a) – (b)). We then apply a pre-filtering to reduce illumination effects and to prevent some local image regions to dominate the energy spectrum. The image I(x, y) is decomposed by a set of Gabor filters. The 2-D Gabor filter is defined as follows:

$$h(x,y) = e^{-\frac{1}{2}\left(\frac{x^2}{\delta_x^2} + \frac{y^2}{\delta_y^2}\right)} e^{-j2\pi(u_0 x + v_0 y)} \quad (1)$$

The parameters (δ_x, δ_y) are the standard deviation of the Gaussian envelope along vertical and horizontal directions; (u_0, v_0) refers to spatial central frequency of Gabor filters. As shown in (Fig. 4 (c)), configurations of Gabor filters contains

[1] Gist feature present a brief observation or a report at the first glance of a outdoor scene that summarizes the quintessential characteristics of an image

4 spatial scales and 8 directions. At each scale (δ_x, δ_y), by passing the image I(x,y) through a Gabor filter h(x,y), we obtain all those components in the image that have their energies concentrated near the spatial frequency point (u_0, v_0). Therefore, the Gist vector is calculated using energy spectrum of 32 responses. We calculated averaging over each grid of 16 x 16 pixels on each response, as shown in (Fig. 4 (d)). Totally, a Gist feature vector is reduced to 512 dimensions.

E. Recognition based on kernel descriptor

Recently, Liefeng Bo *et al.* have proposed a principled way to design rich features to capture various visual attributes (gradient, color, texture) [6]. They have learnt compact features from match kernels via kernel approximation and shown that this method outperforms SIFT and other sophisticated feature learning method. Therefore, we would like to try this method on MHI.

The idea of match kernel is as follows. Normally, given an image, traditional methods represent each image patch by a gradient/ color/ texture vector. This vector is hand-crafted. The work in [6] proposes a kernel view of features. For example, in case of gradient features, the gradient kernel is represented in the following equation (Fig. 5). In this figure, P, Q are two image patches to be compared, u, v are pixels in the image patch. Match kernels defined over various pixel attributes provide a unified way to generate a rich, diverse visual feature set, which has been shown to be very successful to boost recognition accuracy.

Match kernels provide a principled way to measure the similarity of image patches, but evaluating kernels can be computationally expensive when image patches are large.

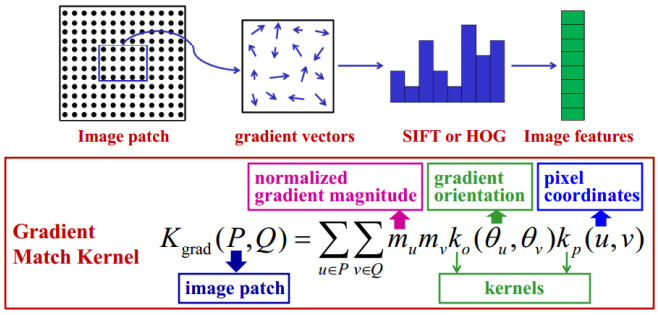

Fig. 5: Gradient Match Kernel

Therefore the authors have proposed a method to reduce as follows. From the match kernel equation, we can derive the feature over image patch:

$$F_{\text{grad}}(P) = \sum_{z\in P} \tilde{m}(z)\phi_o(\tilde{\theta}(z)) \otimes \phi_p(z)$$

Where $\otimes$ is Kronecker product. A straightforward solution to reduce the high dimensional vector is PCA. As the features are presented in view kernels, the authors call KPCA. Finally, for classification, Support Vector Machine (SVM) is used.

III. EXPERIMENTAL RESULTS

A. Database description

The focus of the challenge is on "multiple instance, user independent learning" of gestures, which means learning to recognize gestures from several instances for each category performed by different users, drawn from a gesture vocabulary of 20 categories (Fig. 6, Fig. 8). A gesture vocabulary is a set of unique gestures, generally related to a particular task.

In this challenge we will focus on the recognition of a vocabulary of 20 Italian cultural/anthropological signs. Look inside the dataset, we found that in a hand gesture category, participants do it in a very different manner. This dataset is therefore much more difficult than one-shot learning dataset of the last year.

The challenge has two main phases: development phase and final evaluation phase. Currently, we are in the development phase, the data for final evaluation is released in more than 10 days, so now for evaluation we need to take a set for training and remaining examples for testing from the development data.

At the development phase, we are provided with a large database of 7754 video shots, each contains one hand gesture from 20 gesture categories of Italian signs. Due to the limitation of time, we propose to deal on a sub-set data. Specifically, each hand gesture type have 197 video shots, we take 98 video shots for training and 99 video shots for testing. Totally, we have 3940 video shots, 1960 for training and 1980 for testing.

B. Experimental results

We compare three methods: PCA-KNN, Gist-KNN, KDES-A.

The Fig. 7 shows the recognition rate using PCA-KNN method in both cases: using RGB and D image. In this case, we set K = 5 for KNN. We have varied the value of coefficient number in PCA space (16, 64, 128) and found that 64 gives the best results so we choose this value for latter evaluation. The average of recognition rate is about 21.7% using RGB and 17.7% using depth information. These recognition rates are not high, but as we notices above, the dataset is very challenge: one hand gesture category performed by two persons is highly different. But the category 'Basta' ID = 13 obtain the highest results (RGB: 52%; D: 75.5%). Generally, Using depth information is better than using RGB.

The Fig. 9 shows the recognition rate using GIST-KNN method in both cases: using RGB and D image. In this case, we set also K = 5 for KNN. Different to PCA-KNN, this time, we obtain has the highest recognition rate on the hand gesture 'Vattenne' ID = 1. However, the average of recognition rate is 13%, lower than PCA-KNN that gives an idea that the GIST is very good to represent scene by one frame, but GIST computed on the MHI cannot represent well the video.

The Fig. 10 shows the recognition rate using kernel descriptor in both cases: using RGB and Depth image. In this experiment, we use gradient kernel descriptor. The average accuracy on the depth image (56.4%) is higher than the RGB image (51.3%).

Fig. 6: Example of the first ten gestures

Fig. 8: Example of the last ten gestures

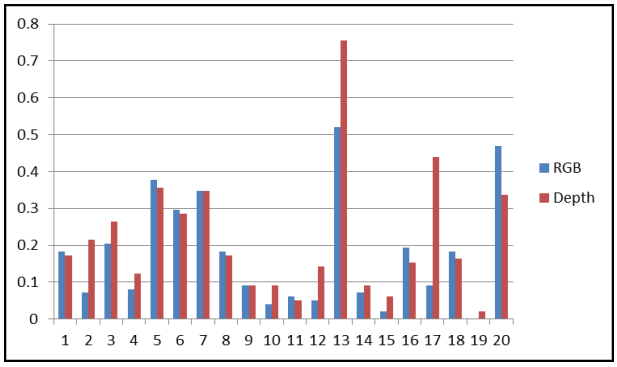

Fig. 7: Recognition rate using PCA-KNN

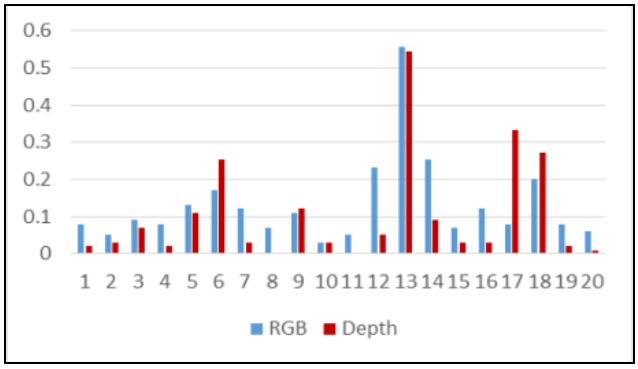

Fig. 9: Recognition rate using GIST-KNN

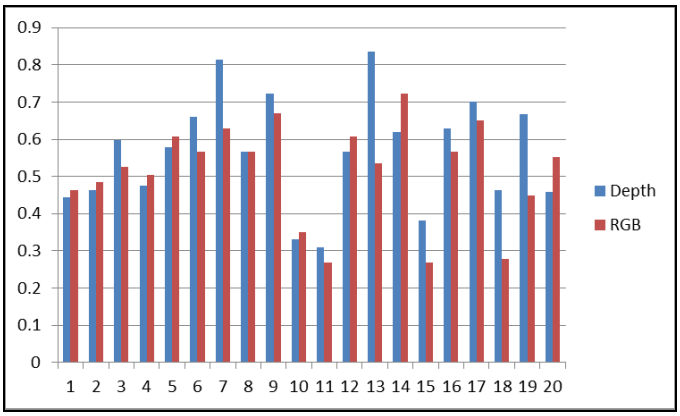

Fig. 10: Recognition rate using KDES-G

IV. CONCLUSION

We have presented our evaluation of three methods on dynamic hand gesture recognition. The first one is based on traditional method that is PCA-KNN. The second one is based on GIST-KNN, very famous for scene categorization. The third one is based on a very new method on kernel descriptor. The experimental results show that the method KDES-G gives the highest recognition rate, the second one is PCA-KNN. In all cases, using depth map gives higher results than using RGB map. This comparative study gives results on a new work research that no results are reported until now. In the future, we will combine both depth and RGB information to improve the recognition rate. We notice that if we based on vision information, the performance cannot be good because by eye we observe very difference between instances in a category. We think that using multimodal information like speech will help to improve the results.

REFERENCES

[1] D. Wu, F. Zhu, and L. Shao, One shot learning gesture recognition from RGBD images, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),2012. p. 7-12.

[2] I. Laptev and T. Lindeberg, Space-time interest points, in ICCV2003.

[3] A. Bobick and J. Davis, The recognition of human movement using temporal templates, in Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing2001.

[4] Quattoni, A. and A.Torralba, Recognizing Indoor Scenes. In Proceeding of the International Conference on Computer Vision and Pattern Recognition, 2009: p. 1-8.

[5] Oliva, A. and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision, 2001. 42(3): p. 145-175.

[6] L. Bo, X. Ren, and D. Fox, Kernel Descriptors for Visual Recognition, in NIPS2010.