Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International conference on Intelligent Human Computer Interaction, IHCI 2014

Hand posture recognition using Kernel Descriptor

Van-Toi NGUYEN[a,b,*], Thi-Lan LE[a], Thanh-Hai TRAN[a], Rémy MULLOT[b], Vincent COURBOULAY[b]

[a]*International Research Institute MICA, HUST-CNRS/UMI-2954-GRENOBLE INP and Hanoi University of Science & Technology, Vietnam*
[b]*L3i Laboratory, the University of La Rochelle, France*

Abstract

In this paper, we propose to investigate the role of a new descriptor named Kernel Descriptor (KDES), recently introduced in [1] for hand posture recognition. As the hand posture has it own the color characteristic, we will examine kernel descriptor in diffident color channels such as HSV, RGB, Lab to find out the most suitable color space for kernel representation of hand posture. We perform extensive experiments on two datasets. The obtained results are promising (97.3% on NUS-2 dataset and 85.0% on our dataset). Thank to the analysis, kernel descriptor is highly recommended for hand posture recognition.

Peer-review under responsibility of the Scientific Committee of IHCI 2014.

Keywords: hand posture recognition, human-robot interaction, kernel descriptor

1. Introduction

Visual interaction provides a natural and effective means for human-system interaction in comparison with different modalities such as audio, tactile. Face and hand are two main parts of the body that are widely used in human-system interaction. While face and facial expression information gives the information about the identity and the emotion/attitude of the person, hand is usually used to provide more information about what people want (commands) in interaction with the system. Hand gesture recognition is a challenging research topic due to the large diversity of hand postures/gestures produced by different people. We distinguish hand posture and hand gesture. A hand posture is the pose of the hand at an instant while a hand gesture is a sequence of hand postures in a duration of time. This means that hand gesture has dynamic aspect.

In our work, we focus on hand posture recognition because in most of human robot interaction applications, hand posture is enough to explain or command something to the robot. Concerning hand posture recognition, the most crucial part is hand posture representation. This part aims at extracting relevant features/descriptors for hand posture. Current hand posture representation methods can be divided into two main categories: implicit and explicit approaches. The works belonging to explicit approach [2] try to locate different hand components such as finger tip and formulate the hand pose based on these components while the second approach [3,4] applies generic visual descriptors

* Corresponding author. Tel.: +84-91-284-7077 ; fax: +84-4-3868-3551.
E-mail address: van-toi.nguyen@mica.edu.vn

Peer-review under responsibility of the Scientific Committee of IHCI 2014.

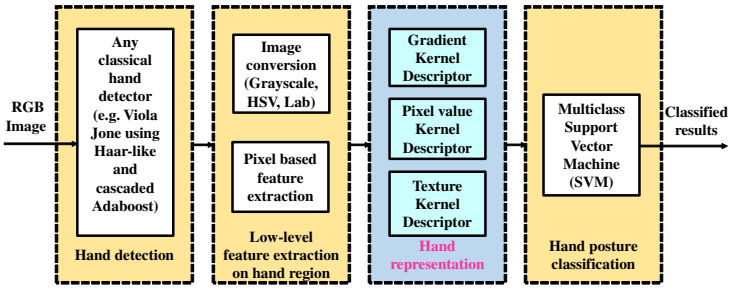


Fig. 1. Our proposed framework for hand posture recognition

such as shape context, color in order to represent in implicit way the hand pose. Recently, Bo et al.[1] has proposed a new descriptor named Kernel descriptor (KDES) for object recognition. This descriptor has been proved to be very robust for several applications. Hence, we propose to apply this descriptor for hand posture recognition.

Our main contributions are following: i) Firstly, kernel descriptor has been evaluated on general object recognition, for the first time, we propose to employ this descriptor for hand posture representation and show how good it is for hand posture recognition; ii) Secondly, as the hand posture has it own the color characteristic, which is different from other objects in the scene, we investigate kernel descriptor in diffident color channels such as HSV, RGB, Lab to find out which color space is the most suitable for kernel representation of hand posture; iii) Finally, we perform an extensive experiments to evaluate the performance of our proposed framework for hand posture recognition using kernel descriptor with two datasets, one is benchmark and one we build ourself to do the comparison.

2. The framework of hand posture recognition

We propose a framework for hand posture recognition based on kernel descriptor as follows. It composes of four main steps (see Fig.1). (1) The step *Hand detection* aims at detecting and localizing hand regions on the image. In our framework, we can apply any hand detection methods. However, in order to make our analysis independent of hand detection results, in the experiment section, we will evaluate the robustness of the hand posture recognition method with perfect hand detection that is done manually. (2) The step *Low-level feature extraction on hand region* takes a hand region as input image, then computes some low-level features such as color conversion, gradient computation, local binary pattern feature extraction, that will be used to build hand-level descriptor. (3) In the step *Hand representation*, we propose to describe hand posture based on kernel descriptor. We will do an extensive analysis on combinations of different kernel descriptors (gradient, color, texture) on different color channels (HSV, Lab, RGB). (4) In the step *Hand posture recognition*, the typical multiclass SVM will be used for hand posture classification. Since the most important step of our framework is hand representation step, in the next section, we will describe it.

3. Hand representation

In this paper, we investigate how kernel descriptor is good for hand posture recognition when it is extracted based on three low level image features that are pixel value, gradient and texture feature on three common color spaces that are RGB, HSV and Lab. We firstly present the definition of the similarity between two image patches beyond match kernel. We then describe how to extract kernel descriptor for each image patch from match kernel.

3.1. Match kernels

We present here an example that is the gradient match kernel. The pixel value match kernel and texture match kernel are formulated in the similar way[1]. The gradient match kernel is constructed from three kernels that are the gradient magnitude kernel $k_{\widetilde{m}}$, the orientation kernel k_o, and the position kernel k_p.

$$K_{gradient}(P,Q) = \sum_{z\in P}\sum_{z'\in Q} k_{\widetilde{m}}(z,z')k_o(\widetilde{\theta}(z),\widetilde{\theta}(z'))k_p(z,z') \quad (1)$$

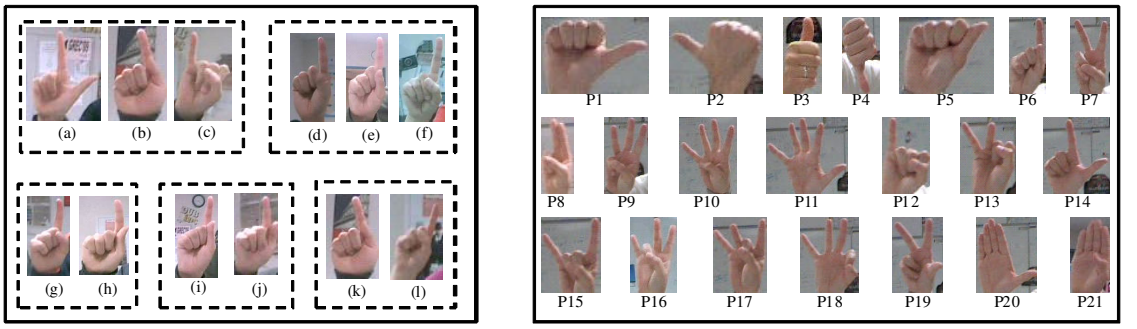


Fig. 2. Our dataset: *Right part*: List of upright postures of right hand in our dataset. *Left part*:Some examples of high inter-class similarity and low intra-class similarity characteristics of our dataset: (a), (b) and (c) are three different hand postures with similar appearance, (d)-(l) images of the same posture under different conditions

where P and Q are patches of two different images we need to measure the similarity. z and z' denote the 2D position of a pixel in the image patch P and Q respectively. The gradient magnitude kernel $k_{\widetilde{m}}$ is defined as the inner product of two normalized gradient magnitude vectors $k_{\widetilde{m}(z)}$ and $k_{\widetilde{m}(z')}$: $k_{\widetilde{m}}(z, z') = k_{\widetilde{m}(z)} k_{\widetilde{m}(z')}$. This inner product is a positive definite kernel. The normalized gradient magnitude $\widetilde{m}(z)$ is defined as following: $\widetilde{m}(z) = \frac{m(z)}{\sqrt{\sum_{z \in P} m(z)^2 + \epsilon_g}}$, where ϵ_g is a small constant. $m(z)$ is magnitude of the image gradient at a pixel z. In Eq.1, the normalized gradient vectors $\widetilde{\theta}(z)$ is defined as following: $\widetilde{\theta}(z) = (sin(\theta(z)), cos(\theta(z)))$, where $\theta(z)$ is orientation of the image gradient at a pixel z. Both the orientation kernel k_o and the position kernel k_p are Gaussian kernels which is of the form: $k(x, x') = \exp(-\gamma \|x - x'\|^2)$. The factor γ will be defined individually for k_o and k_p that are denoted by γ_o and γ_p respectively.

3.2. Extracting kernel descriptors

To extract the compact low dimensional features from match kernels, compact basis vectors need to be generated by learning. With gradient kernel descriptor(Gradient-KDES), let we have a learned set of d_o basis vectors $\{\phi_o(x_1), \phi_o(x_2), ..., \phi_o(x_{d_o})\}$ and a set of d_p basis vectors $\{\phi_p(y_1), \phi_p(y_2), ..., \phi_p(y_{d_p})\}$ considering k_o and k_p kernels respectively. Where x_i are sampled normalized gradient vectors and y_i are normalized $2D$ position of pixels in an image patch. Let we have also ${\alpha^t}_{ij}$ is learned through kernel principal component analysis. Where ${\alpha^t}_{ij}$ are coefficients of the $t - th$ kernel principal component. Then, the gradient kernel descriptor (Eq.1) has the form[1]: $\overline{F}^t_{grad}(P) = \sum_{i=1}^{d_o} \sum_{j=1}^{d_p} \alpha^t_{ij} \{\sum_{z \in P} \widetilde{m}(z) k_o(\widetilde{\theta}(z), x_i) k_p(z, y_j)\}$. The pixel value kernel descriptor (PixelValue-KDES) and the texture kernel descriptor (Texture-KDES) are extracted in a similar way. The combination of the three kernel descriptors (Combination-KDES) is constructed by concatenating the image-level feature vectors. The image-level feature is extracted using efficient match kernels (EMK) that is introduced in[5]. The compact basis vectors then are learned on a general set of images. The learned compact basis vectors, the optimized dimensionality of KPCA and match kernel parameters then be fixed to use in other datasets. We use the set of basis vectors and match kernel parameters that were learned using a subset of ImageNet[1].

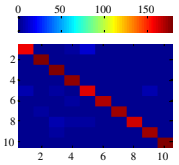
4. Experiments

The aim of these experiments are to evaluate the performance of kernel descriptor for hand posture recognition step as well as to investigate in order to find out the color space that is the most suitable for hand representation. To do this, we use two datasets: one benchmark dataset[4] and one dataset with more number of hand postures that we build ourself.

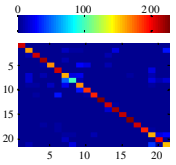
The first dataset is the NUS II dataset introduced in[4]. The second dataset is our dataset collected in the context of human-robot interaction in indoor environment containing cluttered background. This robot stays immobile during interaction with the users. User will stand in front of robot. The distance from robot to user is around 1m to 3m. Based on our study, we define a set of 21 hand postures of up-right hand using in human-robot interaction that is shown in

Descriptor \ Color space \ Dataset	NUS II dataset			Our dataset		
	RGB	HSV	Lab	RGB	HSV	Lab
Gradient-KDES	0.953	0.921	0.951	0.844	0.624	0.624
PixelValue-KDES	0.829	0.803	0.868	0.694	0.725	0.719
Texture-KDES	0.927	0.886	0.903	0.783	0.576	0.688
Combination-KDES	0.961	0.946	**0.973**	**0.850**	0.741	0.823

(a)



(b)



(c)

Fig. 3. (a): The obtained accuracies of our method with two datasets. (b,c) Confusion matrix in the best cases: (b) on NUS II dataset, (c) on our dataset.

the right part of Fig.2. In all postures, the hand points up except the posture #4. The total number of videos in our dataset is 21×10 persons $\times 4$ times $= 840$. Each video has a length of about 4 seconds, a frame rate of 30 fps, and a resolution of 320×240 pixels. The dataset is divided into two parts for training and testing. For each video, every 10 frames we select one sole frame. Our dataset has several challenges for hand detection and hand recognition that are (1) large number of hand postures; (2) high inter-class similarity and low intra-class similarity due to the acquisition condition. The left part of Fig.2 illustrates some challenging examples of our dataset.

We use accuracy to measure the performance of the proposed method. The accuracy is defined by the ratio between the number of correct recognition and the number of test. The average accuracy obtained with two datasets on different descriptors and different color spaces are given in Fig.3(a) while the detail results for each class are illustrated in Fig.3(b,c). This figure shows that our method obtains good results for both datasets. The average accuracy on NUS II dataset is about 10% higher than on our dataset. The reason for this results is that our dataset is more challenging.

The gradient kernel descriptor and texture kernel descriptor obtains the best result on RGB color space while the pixel value kernel descriptor obtains the best result on Lab color space. In over all, the gradient kernel descriptor is often better than the two others. The combination kernel descriptor is lightly better than the best individual descriptor. However, the recognition time of the combination is about three times more expensive than the individual descriptor because the feature extraction time plays an important role in recognition time.

5. Conclusion and future works

In this paper, we have presented a new method that uses kernel descriptor for hand posture recognition. Based on the obtained results, we can see that KDES is prospective descriptor for hand representation of real application. In this case, other descriptors such as shape context can not be employed because it requires a good segmentation while KDES needs only a bounding box of hand region. However, while working with KDES, we have two observations. Firstly, KDES is not invariant to the size of image of object. Secondly, since KDES is computed in both object region and background, the performance of the system depends on the background. These drawbacks should be taken into account in the future work.

Acknowledgements

This research is funded by Hanoi University of Science and Technology under grant number T2014-100.

References

1. Bo, L., Ren, X., Fox, D.. Kernel Descriptors for Visual Recognition. In: *NIPS*. 2010, p. 244–252.
2. Triesch, J., von der Malsburg, C.. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2001;**23**(12):1449 – 1453.
3. Dardas, N.H., Georganas, N.D.. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and Measurement* 2011;**60**(11):3592–3607. doi:10.1109/TIM.2011.2161140.
4. Pisharady, P.K., Vadakkepat, P., Loh, A.P.. Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds. *International Journal of Computer Vision* 2012;doi:10.1007/s11263-012-0560-5.
5. Bo, L., Sminchisescu, C.. Efficient Match Kernel between Sets of Features for Visual Recognition. In: *NIPS*. 2009, p. 135–143.