Nonparametric Gesture Labeling from Multi-modal Data

Ju Yong Chang

Electronics and Telecommunications Research Institute
305-700, Daejeon, Korea
juyong.chang@etri.re.kr

Abstract. We present a new gesture recognition method using multi-modal data. Our approach solves a labeling problem, which means that gesture categories and their temporal ranges are determined at the same time. For that purpose, a generative probabilistic model is formalized and it is constructed by nonparametrically estimating multi-modal densities from a training dataset. In addition to the conventional skeletal joint based features, appearance information near the active hand in the RGB image is exploited to capture the detailed motion of fingers. The estimated log-likelihood function is used as the unary term for our Markov random field (MRF) model. The smoothness term is also incorporated to enforce temporal coherence of our model. The labeling results can then be obtained by the efficient dynamic programming technique. Experimental results demonstrate that our method provides effective gesture labeling results for the large-scale gesture dataset. Our method scores 0.8268 in the mean Jaccard index and is ranked 3rd in the gesture recognition track of the ChaLearn Looking at People (LAP) Challenge in 2014.

Keywords: Gesture recognition, Nonparametric estimation, Multi-modal data

1 Introduction

Human activity recognition is one of the important problems in computer vision and it has various applications such as human-computer interaction, visual surveillance, and intelligent robot. The goal of human activity recognition is to automatically understand human behavior from input data sequence.

Until now, a large amount of research has been conducted for human activity recognition. There are several excellent surveys for RGB image/video based activity recognition [1]. Recently, research based on depth data attracts great attention [2]. The use of depth data enables us to overcome several difficulties of traditional RGB based methods such as appearance variation, illumination change, and loss of 3D information. Moreover, 3D human pose can be efficiently estimated from depth data [3]. According to the recent interesting study [4], such intermediate high-level pose features result in better recognition performance than low/mid-level features such as dense trajectories [5], histograms of oriented gradients [6], and histograms of optical flow [7].

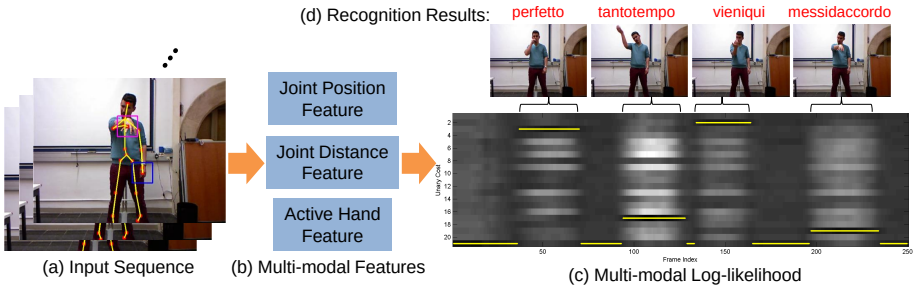

Fig. 1. Overview of the proposed method is illustrated. From the input test sequence (a), multiple features in (b) are computed. They are used to construct the log-likelihood matrix in (c). The x and y-axes denote the frame index and gesture class, respectively. Dark pixels represent the high likelihood region. By using this as the unary term, our MRF model produces the gesture labeling results, that are denoted by the yellow lines. Final recognized gestures are shown in (d).

However, the structure of the estimated 3D pose sometimes lacks enough complexity for certain activity recognition problems. For example, many human gestures contain finger motion, but it is not easy to estimate the 3D pose of the articulated hand model unless close range. In that case, it is necessary to simultaneously consider both the rough information of 3D pose and the detailed appearance of the image. Therefore, in this paper, we focus on human gesture recognition based on multi-modal data, especially using the 3D pose information and the RGB image.

Recently, the Looking at People (LAP) Challenge[1] was held for solving the human gesture recognition problem from multi-modal data [8]. The challenge focused on *multiple instance, user independent spotting* of gestures. Its dataset and goal have the following features. First, it includes a large amount of data, specifically 13,858 gesture instances. Target gestures are defined by twenty Italian cultural and anthropological signs and they are composed of simple atomic motions with both hands. Unlike the setting in many existing gesture/action recognition works, the detection problem should be solved rather than the classification problem where the input video sequence is assumed to be pre-segmented. This means the starting and ending points of the gesture should be also estimated together with its category in this challenge.

In this paper, we present a novel gesture recognition method using multi-modal data. Our proposed method is based on the labeling problem, where a gesture category label should be inferred for every frame in a video. To solve the labeling problem, we basically take a simple classification approach for each frame. Specifically, a multi-class classification problem should be solved to assign every frame in a video one of the multiple labels corresponding to known gesture cat-

[1] http://gesture.chalearn.org

egories. This is a very simple formulation that does not explicitly assume any temporal or hierarchical modeling of human gestures or actions.
The proposed method utilizes following features: skeletal joint position feature, skeletal joint distance feature, and appearance features corresponding to left and right hands. Under the naive Bayes assumption, likelihood functions are independently defined for every feature. Such likelihood functions are nonparametrically constructed from the training data by using kernel density estimation (KDE). For computational efficiency, k-nearest neighbor (kNN) approximation to the exact density estimator is proposed. Constructed likelihood functions are combined to the multi-modal likelihood and this serves as the unary term for our pairwise Markov random field (MRF) model. For enhancing temporal coherence, the smoothness term is additionally incorporated to the MRF model. Final gesture labels can be obtained via 1D MRF inference and this can be efficiently done by dynamic programming. The overview of our method is illustrated in Fig. 1.

2 Related Works

There is extensive literature on action recognition research. In this section, we only review the methods relevant to our approach, especially ones based on depth data. According to the recent surveys [1, 2], action recognition researches can be classified into two categories: sequential approaches and space-time approaches. *Sequential approaches* have traditionally focused on how to model the temporal dynamics of the target actions or gestures. They are usually based on a hidden Markov model (HMM) [9, 10], a conditional random field (CRF) [11], or a graphical model (GM) with more complex structures [12]. These models generally assume the target actions to be represented by the dynamic changes of states and such dynamic patterns are automatically learned from the training data. Our method is also based on the MRF, a kind of GM, and it looks similar to the sequential approaches at first glance. However, our method does not explicitly model the temporal dynamics and the labels of our MRF model represent the gesture categories rather than the intermediate states.
In the dynamic time warping (DTW) based approaches [13, 14], the input test sequence to be recognized is aligned with known sequences of the dataset to produce the alignment-based distances, which are used to determine the action category by finding the best matches. These approaches can be viewed as a nonparametric version of the sequential approaches. Our method is also based on the nonparametric matching process between the input data and the dataset. However, the distance measure in our method is defined as the simple Euclidean distance without the time-consuming alignment step. Therefore, our approach is computationally more efficient than the DTW-based methods and applicable to the large-scale gesture recognition datasets like the LAP multi-modal gesture dataset [8].
Space-time approaches usually take a space-time volume and extract local or global features inside it. The extracted features and the ground-truth action

categories of the training dataset are used to discriminatively learn parametric models by using support vector machines (SVM) [15, 16], structural SVM [17], etc. Such discriminative approaches require a learning step and the performance of the learned model generally depends on the size of the training dataset. Therefore, most of the discriminative approaches suffer from a heavy computational burden for training with the large-scale dataset.

We next investigate nonparametric space-time approaches most relevant to our method. The authors in [18] proposed a nonparametric action recognition method based on the skeletal joint information. In their work, the EigenJoints descriptor is developed and the Naive-Bayes-Nearest-Neighbor (NBNN) classifier is adopted to solve the gesture classification problem. Three features including posture, motion, and offset features are computed from the skeletal joint information. They are then concatenated into one feature vector and principal component analysis (PCA) is used to construct the EigenJoint feature. Their method applies the Naive Bayes (NB) assumption to the EigenJoint features obtained for all frames in a given data sequence. On the other hand, our method utilizes the multi-modal features and assume the NB assumption for the multi-modal features in each frame. While their approach assumes segmented video input for classification, our approach combines the nonparametric model with the MRF to simultaneously solve both classification and segmentation.

In [19], the Moving Pose (MP) descriptor that captures not only the skeletal joint position features but also differential properties like the speed and acceleration of the joints was proposed. The discriminative key frames for each action class are learned from the training dataset and they are used to produce the matching scores between the test sequence and the action classes by using the modified kNN classifier. Moreover, it is applied to all frames according to the sliding window strategy for action detection in unsegmented sequences. However, the labeling problem is locally solved for each frame and temporal coherence between subsequent frames is not enforced. On the other hand, our method adopts the pairwise MRF with the smoothness term, so temporally coherent solutions can be obtained. While the MP descriptor is developed only for the skeletal joint features, our approach can handle the multi-modal data by probabilistic fusion of the nonparametrically estimated likelihood functions.

3 Proposed Method

Suppose we are given a training dataset and each frame in a training sequence is labeled a gesture category $g^{(i)} \in \mathcal{G}$, where i and $\mathcal{G}$ denote the frame index and the set of all gesture labels, respectively. In this paper, each sequence can have multiple gesture categories without overlapping, that is, each frame in a sequence is constrained to be labeled only with one gesture category. Now the objective is to solve the *labeling problem*, where each frame of a test sequence should be assigned a gesture category label.

3.1 Generative Probabilistic Model

We basically approach to the gesture labeling problem by solving the classification problem, where temporal positions or dynamics are not considered at all and the gesture should be independently classified for each frame. This is an extremely simple assumption compared to the general HMM or CRF based approaches that usually introduce the intermediate states and model gestures by the temporal dynamics of the states. Now the generative probabilistic model for gesture classification can be simply formalized as follows. Hidden random variables G generate M multi-modal observations $X_j, j \in \{1, \ldots, M\}$. Here the observed feature X_j is computed from the multi-modal data in the several frames near the current frame. Under the naive Bayes assumption, the multi-modal features are conditionally independent of each other given the gesture category. Therefore, the multi-modal likelihood can be defined by

$$p(X_1, \ldots, X_M|G) = p(X_1|G) \cdots p(X_M|G). \tag{1}$$

Now we present the multi-modal features and how to estimate their corresponding likelihoods.

3.2 Multi-modal Features

In this paper, the skeletal joint data and RGB images are assumed to be the multi-modal input to our proposed method. It is well known that the skeletal joint features can be efficiently and robustly estimated from the depth image [3]. From the skeletal joints, we only consider K joints belonging to the upper body. Let $x_j, j = 1, \ldots, K$ denote the 3D coordinates of such joints. We then define the normalized joint coordinates $\bar{x}_j, j = 1, \ldots, K$ by taking the differences between x_j and the reference joint x_p, that is assumed to be the *neck* joint in this paper. To increase the discriminability, we concatenate the normalized joint coordinates from L_P frames near the current frame to construct the *skeletal joint position feature* $\mathbf{x}_\mathrm{P}$. The resultant $\mathbf{x}_\mathrm{P}$ is a $L_\mathrm{P} \cdot 3 \cdot K$ dimensional vector and it holistically describes the motion dynamics of the upper body near the current frame.

Despite the normalization process, the skeletal joint position feature is not viewpoint invariant. As a viewpoint invariant feature, we utilize the Euclidean distance $\|x_j - x_k\|$ between joint j and k. The *skeletal joint distance feature* $\mathbf{x}_\mathrm{D}$ is then defined by concatenating all such distances for L_D frames. Note that the dimensionality of $\mathbf{x}_\mathrm{D}$ is $L_\mathrm{D} \cdot \frac{K(K-1)}{2}$. This is a kind of the relational pose feature [20], describing geometric relations between specific joints in a short sequence of frames.

We additionally consider the RGB image to exploit the details not captured by the skeletal joint features. For that purpose, the 3D joints of left and right hands are first projected to the RGB image. Histogram of oriented gradients (HOG) descriptors are then computed for the windows centered on the projected points. We concatenate the HOG descriptors of L_L frames near the current frame to construct our *appearance feature* $\mathbf{x}_\mathrm{L}$ for the left hand. The appearance feature for

the right hand $\mathbf{x}_{\mathrm{R}}$ is similarly defined from the HOG descriptors corresponding to the right hand.
Because our features are constructed from several frames, their dimensionality is generally very high, especially for the appearance features. Therefore, we use the PCA to reduce the computational complexity of our method. We also apply the standardization process to compensate the different scales of the multi-modal features. As a result, each multi-modal feature will have zero-mean and unit-variance.

3.3 Active Hand Approach

In general, gestures including the motion of hands often express the information with just one hand and which hand to use is not important. For gestures based on both hands, their motions are usually similar to each other. Based on these observations, we propose to select the main hand and to use its appearance feature for gesture representation. For that purpose, we introduce a new deterministic variable $a^{(i)}$ for each frame i:

$$a^{(i)} = \begin{cases} 0, \text{ if } \|x_l^{(i+1)} - x_l^{(i)}\| > \|x_r^{(i+1)} - x_r^{(i)}\|; \\ 1, \text{ otherwise,} \end{cases} \tag{2}$$

where $x_l^{(i)}$ and $x_r^{(i)}$ denote the 3D joint coordinates of left and right hands, respectively. The variable $a^{(i)}$ can be intuitively understood as an indicator of which hand is more active at i-th frame. Now our hypothesis is that the active hand is the main hand and using only the feature of the main hand is helpful for gesture classification. We finally define the *appearance feature for the active hand* $\mathbf{x}_{\mathrm{A}}$ by $\mathbf{x}_{\mathrm{L}}$ if the left hand is active ($a = 0$) and by $\mathbf{x}_{\mathrm{R}}$ if the right hand is active ($a = 1$). In this paper, this active hand feature $\mathbf{x}_{\mathrm{A}}$ is adopted instead of the left and right hand features $\mathbf{x}_{\mathrm{L}}$ and $\mathbf{x}_{\mathrm{R}}$.

3.4 Nonparametric Estimation of Multi-modal Likelihood

Now we present how to estimate the likelihood function for each feature from the training dataset. Let $\mathbf{x}_1^g, \ldots \mathbf{x}_N^g$ denote all the features labeled a gesture category class g from all the training sequences. Then the kernel density estimator of the likelihood function is:

$$\hat{p}(\mathbf{x}|g) = \frac{1}{N} \sum_{j=1}^{N} K(\mathbf{x} - \mathbf{x}_j^g), \tag{3}$$

where $K(\mathbf{x})$ is the kernel function, which should be non-negative and integrate to one. In this paper, the spherical Gaussian function is used for the kernel function $K(\mathbf{x}) = (2\pi)^{-D/2}\sigma^{-D} \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}\|^2)$, where D and σ denote the dimensionality of the feature vector and the bandwidth parameter, respectively.
In general, N (i.e., the number of training samples belonging to each gesture class) is very large, so computing the likelihood in (3) is computationally very expensive. Therefore we approximate it by considering only the largest term in

the summation (3). Because the Gaussian kernel is assumed, this term corresponds to the nearest neighbor of the feature vector $\mathbf{x}$ within $\mathbf{x}_1^g, \ldots \mathbf{x}_N^g$, and the likelihood function (3) can be rewritten as:

$$p^{\mathrm{NN}}(\mathbf{x}|g) \propto \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_{\mathrm{NN}}^g\|^2), \tag{4}$$

where $\mathbf{x}_{\mathrm{NN}}^g$ denotes the nearest neighbor vector. This nearest neighbor vector can be efficiently found by using the randomized kd-trees [21]. In general, we can consider the k-nearest neighbors ($k \geq 2$), but empirically this improves the performance very little.

We apply the above estimation process to all multi-modal features, and obtain their corresponding approximate likelihoods. They are then combined to the multi-modal likelihood and its negative log-likelihood can be written as

$$L(\mathbf{x}|g) = \sum_{j=1}^{M} \frac{1}{2\sigma_j^2}\|\mathbf{x}_j - \mathbf{x}_{j,\mathrm{NN}}^g\|^2, \tag{5}$$

where σ_j is the bandwidth for the j-th multi-modal feature and $\mathbf{x}_{j,\mathrm{NN}}^g$ denotes the nearest neighbor of the multi-modal feature $\mathbf{x}_j$ within the training samples of the gesture class g. Note that the bandwidth parameters $\sigma_j, j \in \{1, \ldots, M\}$ control the relative importance between the multi-modal features. They can be decided based on several approaches such as maximum likelihood criterion [22], discriminative method with the hinge loss [23], etc. In this paper, we simply use the cross-validation. This requires us to search a 2D parameter space of $(\frac{\sigma_{\mathrm{D}}}{\sigma_{\mathrm{P}}}, \frac{\sigma_{\mathrm{A}}}{\sigma_{\mathrm{P}}})$, which is feasible.

3.5 MRF Model with Temporal Coherence

Now let us assume that the test sequence is given and its multi-modal features are $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_M^{(i)})$, $i = 1, \ldots, T$, where T is the length of the test sequence. We can then locally perform gesture labeling for the test sequence by using the negative log-likelihood in (5). Specifically, for each frame i of the test sequence, the optimal gesture class $g^{(i)*}$ can be found by minimizing the negative log-likelihood:

$$g^{(i)*} = \arg\min_{g^{(i)}} L(\mathbf{x}^{(i)}|g^{(i)}). \tag{6}$$

However, this locally optimized solution may lack the temporal coherence. Therefore, we formulate the following MRF model to enhance the temporal coherence of the solution:

$$E(\mathbf{g}) = \sum_{i=1}^{T} E_{\mathrm{unary}}(\mathbf{x}^{(i)}, g^{(i)}) + \lambda \sum_{i=1}^{T-1} E_{\mathrm{pairwise}}(g^{(i)}, g^{(i+1)}), \tag{7}$$

where $\mathbf{g} = (g^{(1)}, \ldots, g^{(T)})$ denotes the gesture label vector. The unary term is defined as the negative log-likelihood ratio:

$$E_{\mathrm{unary}}(\mathbf{x}^{(i)}, g^{(i)}) = L(\mathbf{x}^{(i)}|g^{(i)}) - \min_{g \in \bar{g}^{(i)}} L(\mathbf{x}^{(i)}|g), \tag{8}$$

where $\bar{g}^{(i)}$ denotes the set of all gesture classes excluding $g^{(i)}$. This slightly improves the performance rather than using the negative log-likelihood. And the pairwise term is defined as the simple smoothness constraint:

$$E_{\text{pairwise}}(g^{(i)}, g^{(i+1)}) = \begin{cases} 0, \text{ if } g^{(i)} = g^{(i+1)}; \\ 1, \text{ otherwise.} \end{cases} \tag{9}$$

The parameter λ controls the strength of the smoothness constraint and it is determined by the cross-validation. Now the final gesture label vector $\mathbf{g}$ can be obtained by minimizing the MRF energy in (7). Because our model is 1D MRF, its optimal solution can be very efficiently computed by using the dynamic programming.

4 Experimental Results

4.1 Dataset and Evaluation Metric

To evaluate the performance of the proposed gesture labeling method, we use the gesture dataset [8] introduced in ChaLearn LAP Challenge. It is composed of total 940 sequences (470 training, 230 validation, and 240 test sequences) and each sequence contains RGB, depth data, skeleton information extracted from the depth data by [3], and manually annotated gesture labels. Target gestures are twenty Italian cultural/anthropological signs performed by many subjects. Specifically, there are total 13,858 gesture instances (7,754 training, 3,362 validation, and 2,742 test instances) and this is one of the largest-known datasets for gesture recognition.

Let $A_{(s,n)}$ and $B_{(s,n)}$ denote the ground-truth of gesture n at sequence s and its prediction result, where both $A_{(s,n)}$ and $B_{(s,n)}$ are sets including frames at which the n-th gesture is being performed in the s-th sequence. The Jaccard index can then be defined as

$$J_{(s,n)} = \frac{|A_{(s,n)} \cap B_{(s,n)}|}{|A_{(s,n)} \cup B_{(s,n)}|}, \tag{10}$$

which represents the similarity between two sets. The Jaccard index $J_{(s,n)}$ is averaged over all gesture classes and all sequences to produce the *mean Jaccard index*. We use this mean Jaccard index as the main evaluation criterion. We also compute the *precision* and *recall* to evaluate the detection performance of our method. For that purpose, we need to judge whether the detected gesture interval is the true/false positive. Similarly to the object detection research [24], the detection result is considered to be correct if the overlap ratio r between the ground-truth interval I_{gt} and the predicted interval I_{p} exceeds 0.5:

$$r = \frac{\text{length}(I_{\text{gt}} \cap I_{\text{p}})}{\text{length}(I_{\text{gt}} \cup I_{\text{p}})}, \tag{11}$$

where $I_{\text{gt}} \cap I_{\text{p}}$ represents the intersection of the ground-truth and predicted intervals and $I_{\text{gt}} \cup I_{\text{p}}$ their union.

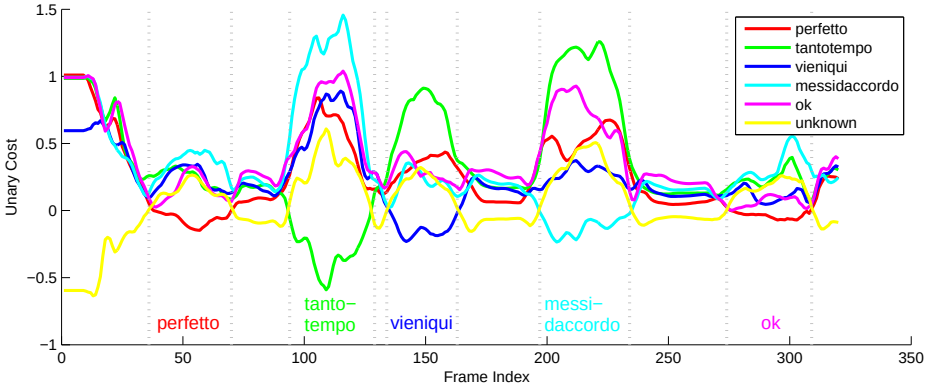

Fig. 2. Unary costs for one of the test sequences (sample index 701) are illustrated in this figure. Colored texts and gray dotted lines denote the ground-truth gesture classes and their starting/ending frames, respectively.

4.2 Implementation Details

Each frame of the dataset sequence is labeled one of the twenty gesture categories, but there are also many frames containing no meaningful gestures. We simply regard it as the twenty-first gesture category in our gesture labeling framework. However, note that it is excluded from computing the evaluation metrics such as the Jaccard index, precision, and recall.

For the skeletal joint position and distance features, $K = 10$ joints in the upper body are utilized. The RGB images of both hands are resized to 128×128 images, and their HOG descriptors are computed with 16×16 cell size and 9 orientations. The size of the temporal window for constructing our features are $L_{\mathrm{P}} = 14$, $L_{\mathrm{D}} = 14$, and $L_{\mathrm{A}} = 2$, respectively. We then apply the PCA to each feature for dimensionality reduction. The variance thresholds of the PCA are 0.99 for the skeletal joint position and distance features, and 0.85 for the appearance feature of the active hand. Remaining parameters are the bandwidth parameters σ_{P}, σ_{D}, σ_{A}, and the smoothness parameter λ. They are determined by the cross-validation. Finally, the open source VLFeat library [25] is used to compute the HOG descriptors and to perform the fast nearest neighbor search with the randomized kd-trees.

4.3 Performance Analysis

From now, we present the evaluation results of the proposed gesture labeling method. First, to investigate the feasibility of our method, we only consider the skeletal joint position feature. Fig. 2 shows the unary costs for a test sample with the annotated ground-truth. We can see that the unary costs provide the strong cues for gesture labeling. The value of the unary cost (i.e., negative log-likelihood ratio) for the ground-truth gesture class is lowest among all the gesture classes

Feature	mJ	mP	mR
Joint Pos. (Local)	0.7547	0.7891	0.8556
Joint Position	0.7816	0.8653	0.8754
Joint Distance	0.7536	0.8403	0.8496
Left Hand	0.3613	0.6505	0.4421
Right Hand	0.6412	0.8270	0.7396
Both Hands	0.7136	0.9013	0.8494
Active Hand	0.7504	0.8885	0.8822

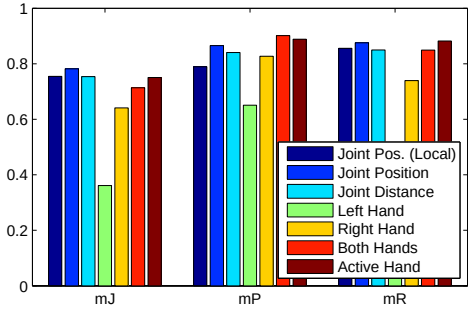

Fig. 3. Performance of the proposed method based on a single feature is illustrated. mJ, mP, and mR denote the mean Jaccard index, mean precision, and mean recall, respectively.

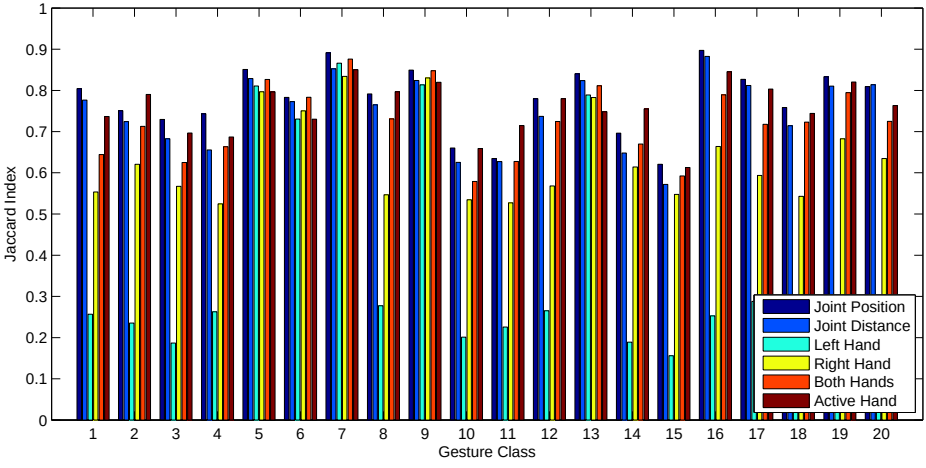

Fig. 4. Jaccard index scores are illustrated for each gesture category.

except the *ok* gesture in Fig. 2. Therefore, the proposed method can produce satisfactory results without considering the coherence between the neighboring frames. It is well illustrated in Fig. 3, where the evaluation results of our method with and without the smoothness term can be found. The mean Jaccard index of our results based on the local approach (6) is 0.7547. We then adopt the smoothness term in (9) and globally optimize the MRF energy in (7), which results in 0.7816. Note that the local approach achieves a reasonable performance and the global version further improves it.

We next examine the performance of the proposed method based on a single feature. Fig. 3 shows the evaluation results corresponding to the various features introduced in Section 3.2 and 3.3. Note that the right hand based appearance feature results in better performance than the left hand based one. This is because the right hand is more frequently used in many gesture instances of the

(a) *ok* (b) *cheduepalle* (c) *chevuoi*

(d) *daccordo* (e) *combinato* (f) *basta*

Fig. 5. The pose of fingers plays an important role in recognizing the *ok* gesture in (a). The gestures based on the motions of both hands are illustrated in (b)-(f).

dataset. Both hands are then simultaneously utilized and this outperforms the single hand cases as expected. Finally, the active hand feature defined in Section 3.3 is evaluated, which produces better results than using the both hands. Note that only half the amount of information is required to represent the active hand feature compared to the both hands.

For detailed analysis, the Jaccard index scores are computed for each gesture category and they are illustrated in Fig. 4, where the numbers of x-axis denote the following twenty gesture classes: *vattene*, *vieniqui*, *perfetto*, *furbo*, *cheduepalle*, *chevuoi*, *daccordo*, *seipazzo*, *combinato*, *freganiente*, *ok*, *cosatifarei*, *basta*, *prendere*, *noncenepiu*, *fame*, *tantotempo*, *buonissimo*, *messidaccordo*, and *sonostufo*. We can see that the active hand based approach significantly outperforms the others especially for the gesture category 11, i.e., *ok*. This is reasonable because the specific configuration of fingers is an important characteristic to distinguish the *ok* from the other gestures. It is also noticeable that the left hand based method achieves comparable results with the right hand for the gesture classes 5, 6, 7, 9, 13, i.e., *cheduepalle*, *chevuoi*, *daccordo*, *combinato*, and *basta*. All these gestures are composed of the same motions of both hands. See Fig. 5 that illustrates the above mentioned gestures.

We then investigate the performance of our multi-modal approach. Let us consider the joint position $\mathbf{x}_{\mathrm{P}}$, joint distance $\mathbf{x}_{\mathrm{D}}$, and active hand $\mathbf{x}_{\mathrm{A}}$ based features. To exploit the multiple features, their bandwidth parameters σ_{P}, σ_{D}, and σ_{A} should be decided. We set σ_{P} to 1.0 and then optimize σ_{D} and σ_{A} by the brute-force search. The smoothness parameter λ is similarly determined. Fig. 6 illustrates the results of this cross-validation process with the validation dataset.

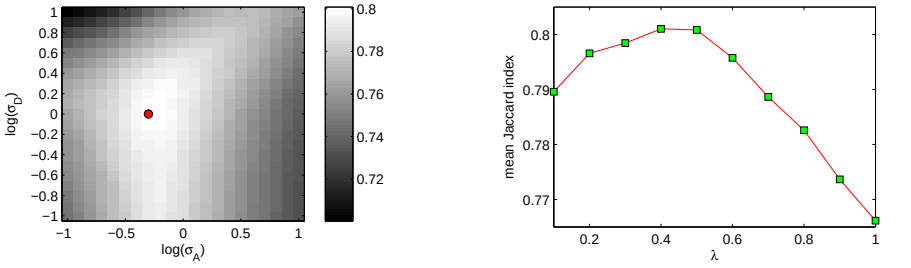

Fig. 6. The bandwidth parameters and the smoothness coefficient are determined by the simple brute-force search.

Feature	mJ	mP	mR
P	0.7816	0.8653	0.8754
P+D	0.7948	0.8962	0.8824
P+L+R	0.8110	0.9254	0.9157
P+A	0.8244	0.8950	0.9191
P+D+A (NoCV)	0.8247	**0.9258**	**0.9204**
P+D+A	**0.8268**	0.9199	0.9158

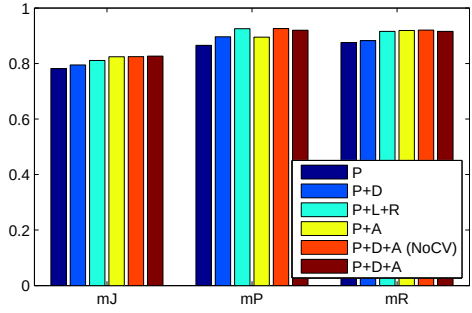

Fig. 7. Performance of the proposed method based on multiple features is illustrated. P, D, L, R, and A denote the joint position, joint distance, left hand, right hand, and active hand based features, respectively. NoCV denotes that cross-validation is not used and all bandwidth parameters are equally set to 1.0.

Fig. 7 shows that the mean Jaccard index score is slightly improved by optimizing the bandwidth parameters. Note that the simple feature combination with equal bandwidth produces the comparable result 0.8247 with the optimal case 0.8268 thanks to the standardization process between the multi-modal features. Fig. 7 and 8 show the evaluation results by using various combinations of multiple features. We can see that the use of multiple features significantly improves the recognition performance. Specifically, by using the joint position, joint distance, and active hand features together, the proposed method scores 0.8268 in the mean Jaccard index and this result is a 3rd place in the gesture recognition track of the ChaLearn LAP Challenge. The top 10 results of the challenge are reported in Table 1.

We finally examine the computational complexity of the proposed method. Our algorithm can be roughly divided into two parts: (1) the training process including feature pre-processing and kd-tree construction, (2) the testing process including nearest neighbor search with the kd-trees and MRF optimization by dynamic programming. The training part uses the training and validation datasets,

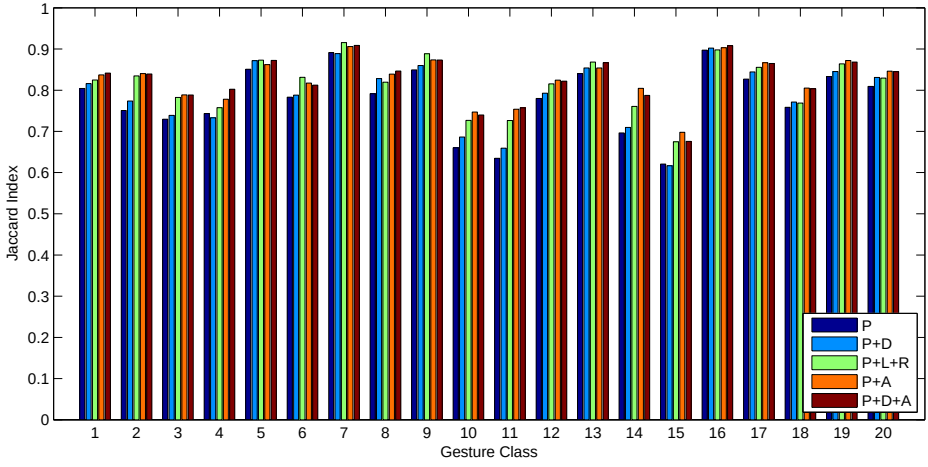

Fig. 8. Jaccard index scores are illustrated for each gesture category.

whereas the testing part runs on the test dataset. The run-time results of them are 1191.7 and 474.1 seconds respectively, as measured on a 12 core CPU machine. Specifically, it takes 1.4 milliseconds per frame to perform the gesture labeling of a test sequence, which shows the efficiency of our approach.

Table 1. Results of ChaLearn LAP Challenge (track 3) are illustrated.

Rank	Team	Score	Rank	Team	Score
1	Neverova et al. [26]	0.8500	6	Wu [27]	0.7873
2	Monnier et al. [28]	0.8339	7	Camgoz et al. [29]	0.7466
3	**Ours**	**0.8268**	8	Evangelidis et al. [30]	0.7454
4	Peng et al. [31]	0.7919	9	Undisclosed authors	0.6888
5	Pigou et al. [32]	0.7888	10	Chen et al. [33]	0.6490

5 Conclusions

We have proposed a novel gesture recognition method based on the nonparametric density estimation of the multi-modal features. Our approach can produce gesture category labels for all frames of the test sequence, which allows not only gesture classification but also accurate localization. Experimental results demonstrate that the proposed method achieves a convincing performance in terms of the mean Jaccard index criterion. In our method, the bandwidth parameters and the smoothness coefficient are determined by the simple cross-validation. The computational complexity of this brute-force search grows exponentially

with the number of features. In the future, we aim to estimate these parameters by using a more sophisticated learning process.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP. [2014(APP0120130417001), Development of High Accuracy Mobile and Omnidirectional Multi-user Gesture Recognition Technology for Interaction with Content]

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. **43**(3) (2011) 16:1–16:43
2. Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J.: A survey on human motion analysis from depth data. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. Springer Berlin Heidelberg (2013) 149–187
3. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. (2011) 1297–1304
4. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. (2013) 3192–3199
5. Wang, H., Klser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1) (2013) 60–79
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. Volume 1. (2005) 886–893
7. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
8. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: ECCV Workshops. (2014)
9. Jalal, A., Uddin, M.Z., Kim, J.T., Kim, T.S.: Recognition of human home activities via depth silhouettes and r transformation for smart homes. Indoor and Built Environment (2011) 1–7
10. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: CVPR Workshops. (2012) 20–27
11. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: ICCV. Volume 2. (2005) 1808–1815
12. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: IEEE International Conference on Robotics and Automation (ICRA). (2012) 842–849
13. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: ICCV. (2009) 444–451
14. Wang, J., Wu, Y.: Learning maximum margin temporal warping for action recognition. In: ICCV. (2013) 2688–2695
15. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. (2012) 1290–1297

16. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: CVPR. (2013) 716–723
17. Wei, P., Zheng, N., Zhao, Y., Zhu, S.C.: Concurrent action detection with structural prediction. In: ICCV. (2013) 3136–3143
18. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In: CVPR Workshops. (2012) 14–19
19. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: ICCV. (2013) 2752–2759
20. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. ACM Transactions on Graphics (TOG) **24**(3) (2005) 677–685
21. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application (VISAPP). (2009) 331–340
22. Murillo, J.M.L., Rodríguez, A.A.: Algorithms for gaussian bandwidth selection in kernel density estimators. Neural Inf. Proc. Systems (2008)
23. Behmo, R., Marcombes, P., Dalalyan, A., Prinet, V.: Towards optimal naive bayes nearest neighbor. In: ECCV. (2010) 171–184
24. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2) (2010) 303–338
25. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)
26. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: ECCV Workshops. (2014)
27. Wu, D.: Deep dynamic neural networks for gesture segmentation and recognition. In: ECCV Workshops. (2014)
28. Monnier, C., German, S., Ost, A.: A multi-scale boosted detector for efficient and robust gesture recognition. In: ECCV Workshops. (2014)
29. Camgoz, N.C., Kindiroglu, A.A., Akarun, L.: Gesture recognition using template based random forest classifiers. In: ECCV Workshops. (2014)
30. Evangelidis, G., Singh, G., Horaud, R.: Continuous gesture recognition from articulated poses. In: ECCV Workshops. (2014)
31. Peng, X., Wang, L., Cai, Z.: Action and gesture temporal spotting with super vector representation. In: ECCV Workshops. (2014)
32. Pigou, L., Dieleman, S., Kindermans, P.J.: Sign language recognition using convolutional neural networks. In: ECCV Workshops. (2014)
33. Chen, G., Clarke, D., Giuliani, M., Weikersdorfer, D., Knoll, A.: Multi-modality gesture detection and recognition with un-supervision, randomization and discrimination. In: ECCV Workshops. (2014)