This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

Nonparametric Feature Matching based Conditional Random Fields for Gesture Recognition from Multi-modal Video

Ju Yong Chang, *Member, IEEE*

Abstract—We present a new gesture recognition method that is based on the conditional random field (CRF) model using multiple feature matching. Our approach solves the labeling problem, determining gesture categories and their temporal ranges at the same time. A generative probabilistic model is formalized and probability densities are nonparametrically estimated by matching input features with a training dataset. In addition to the conventional skeletal joint-based features, the appearance information near the active hand in an RGB image is exploited to capture the detailed motion of fingers. The estimated likelihood function is then used as the unary term for our CRF model. The smoothness term is also incorporated to enforce the temporal coherence of our solution. Frame-wise recognition results can then be obtained by applying an efficient dynamic programming technique. To estimate the parameters of the proposed CRF model, we incorporate the structured support vector machine (SSVM) framework that can perform efficient structured learning by using large-scale datasets. Experimental results demonstrate that our method provides effective gesture recognition results for challenging real gesture datasets. By scoring 0.8563 in the mean Jaccard index, our method has obtained the state-of-the-art results for the gesture recognition track of the 2014 ChaLearn Looking at People (LAP) Challenge.

Index Terms—Gesture recognition, conditional random field, nonparametric estimation, structured learning

1 Introduction

Human activity recognition is one of the most important problems in computer vision that provides various applications, such as human-computer interaction, visual surveillance, image/video retrieval, and intelligent robots. The goal of human activity recognition is to automatically understand human behaviors from input data.

Many studies have investigated human activity recognition until now. Several excellent surveys have also been conducted for RGB image/video-based activity recognition [1], [2]. Studies that are based on depth data have recently attracted an increasing amount of attention [3]. Such data enable researchers 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 by using depth data [4], [5]. It is well known that humans can recognize human activities on a few body joints [6]. In a recent study [7], Jhuang et al. has documented that for their dataset, where the full body is visible, such intermediate high-level pose features result in better recognition performance than low/mid-level features, such as dense trajectories [8], histograms of oriented gradients (HOG) [9], histograms of optical flow [10], and motion boundary histograms [11].

However, the structure of the estimated 3D pose is often insufficiently complex for certain activity recognition problems. For example, many human gestures involve finger motions, but estimating the 3D pose of the articulated hand model is difficult unless the distance between the camera and the hand is sufficiently short. In that case, both the rough information of the 3D pose and the detailed appearance of the image must be considered simultaneously. Therefore, we investigate human gesture recognition by using multi-modal data, especially the 3D pose information and the RGB image.

The Looking at People (LAP) Challenge[1] was recently conducted for solving the human gesture recognition problem from multi-modal data [12]. The challenge focuses on a *multiple-instance, user-independent recognition* of gestures. The dataset and goal of such challenge include the following features. First, the LAP dataset includes a large amount of data, specifically 13,858 gesture instances. Target gestures are defined by 20 Italian cultural and anthropological signs and are composed of simple atomic motions with both hands. Unlike in most existing gesture/action recognition studies, the detection problem should be solved rather than the classification problem where the input video sequence is assumed to be pre-segmented. Therefore, the starting and ending points of the gestures should be estimated simultaneously with their categories in this challenge.

In this paper, we address the problem of gesture recognition from multi-modal video data. For that purpose, we formulate the labeling problem, where a gesture category label must be inferred for every frame in a video. The formulation of the labeling problem has been widely used in computer vision, such as stereo correspondences [13], single view reconstruction [14], and image segmentation [15]. Our proposed method for solving the labeling problem is

• *J. Y. Chang is with the SW-Content Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon 305-700, Republic of Korea.*
E-mail: juyong.chang@etri.re.kr

Manuscript received x x, x; revised x x, x.

1. http://gesture.chalearn.org/2014-looking-at-people-challenge/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

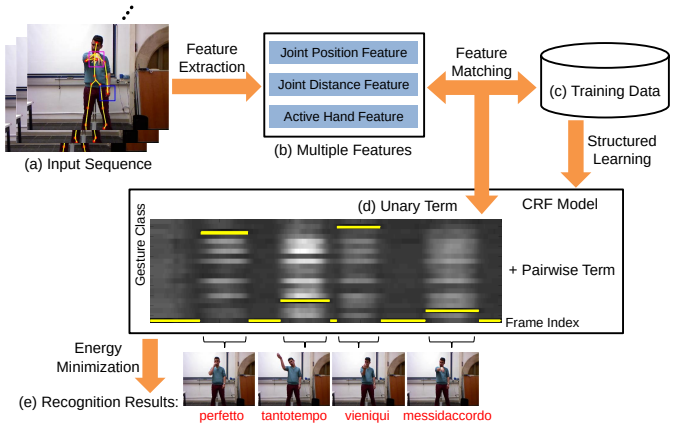


Fig. 1. Overview of the proposed method is illustrated. From an input test sequence (a), multiple features are extracted as in (b). They are matched to training data (c), resulting in a log-likelihood matrix in (d), which is used as the unary term of the proposed CRF model. For parameter estimation of our CRF model, structured learning is performed based on the SSVM framework. In the log-likelihood matrix, the x- and y-axes denote the frame index and the gesture class, respectively. Dark pixels represent high likelihood regions. By minimizing the energy function defined by the CRF model, we can obtain frame-wise recognition results, that are denoted by yellow lines. Final recognized gestures are shown in (e).

inspired by the success of recent nonparametric semantic segmentation methods [16], [17], where segmentation of an image into regions and categorization of all the image pixels are simultaneously performed. Thanks to the nonparametric nature, they can easily adopt new training data and even new object categories without additional training. And they are conceptually simple and easy to implement. These nonparametric approaches are usually based on simple feature matching and co-occurrence statistics between neighboring (super)pixels. Conditional random field (CRF) models provide a natural framework for combining them and efficient inference algorithms. Similarly to such methods, we propose a nonparametric feature matching based CRF model for gesture recognition and its learning strategy.

The proposed method extracts several features from the input data, namely, skeletal joint position feature, skeletal joint distance feature, and appearance features that correspond to the left and right hands. Under the naive Bayes assumption, likelihood functions are independently defined for every feature type. Such likelihood functions are nonparametrically constructed from the training data via kernel density estimation. The k-nearest neighbor approximation to the exact density estimator is adopted for computational efficiency. The constructed likelihood functions are combined to serve as the unary term for our CRF model. To enhance temporal coherence, the smoothness constraint is incorporated as the pairwise term into the CRF model. Final gesture labels can be obtained via 1D inference, which can be efficiently performed by dynamic programming. To train the proposed method with a small number of parameters, we can simply use brute-force search. However, it cannot be applied for more complex models with many parameters. Therefore, we apply the structured learning framework according to the structured support vector machine (SSVM) to the parameter optimization of the proposed CRF model.

Fig. 1 presents an overview of our method.

The main contributions of our work are as follows:

- In this paper, the novel CRF model for gesture recognition is proposed, in which the gesture classes are estimated for all frames to localize the output gestures automatically in the unconstrained input video. Unlike the previous CRF model for gesture recognition [18], the unary term in our CRF model is formalized by applying the generative probabilistic approach and is defined in a nonparametric manner via simple nearest neighbor feature matching. In the proposed gesture recognition framework, different cues from the multiple features can be probabilistically fused into our unary potential. Our nonparametric unary term can naturally handle the multimodality of the feature distribution and the nonlinear decision boundaries.
- We present a novel learning strategy for estimating the parameters of the proposed CRF model. We incorporate the SSVM framework that can be applied to large-scale datasets, such as the LAP Challenge dataset [12]. Unlike the maximum likelihood parameter estimation that easily becomes intractable as the problem size increases, the SSVM can efficiently handle a huge number of linear constraints in its formulation. We empirically prove that the conventional Hamming loss works efficiently for the SSVM framework under the Jaccard index evaluation metric in our experiments.
- We have conducted thorough experiments and find that our approach achieves a state-of-the-art recognition performance. Especially for the LAP Challenge dataset [12], the proposed method scores 0.8563 in the mean Jaccard index, which is considered the best result in the gesture recognition track of the 2014 ChaLearn Looking at People Challenge.

This paper is organized as follows. Section 2 introduces related works, Section 3 presents the proposed model, Section 4 describes its learning method, Section 5 outlines the experimental results, and Section 6 provides the concluding remarks.

2 Related Works

Action recognition has been extensively investigated in the literature. In this section, we only review those studies that are relevant to our approach.

Recent surveys [1], [3] have classified action recognition studies into two categories, namely, sequential approaches and space-time approaches. *Sequential approaches* have traditionally focused on modeling the temporal dynamics of the target actions. These studies are usually based on a hidden Markov model (HMM) [19], [20], [21], a conditional random field [22], or a graphical model (GM) with highly complex structures [23]. These models generally assume the target actions are to be represented by the dynamic changes of states and that such dynamic patterns are automatically learned from the training data. Although based on CRF, our method does not explicitly model temporal dynamics, and its labels represent gesture categories rather than intermediate states.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

In those studies that are based on dynamic time warping [24], [25], [26], the input test sequence to be recognized is aligned with the known sequences of the dataset to produce 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 training dataset. However, the distance measure in the proposed method is defined as the simple Euclidean distance, which makes our approach applicable to large-scale gesture recognition datasets.

Space-time approaches usually take a space-time volume and then extract its local or global features. The extracted features and the ground-truth action categories of the training dataset are used to discriminatively learn the parametric models by using support vector machines (SVM) [27], [28], [29] and structured SVM [30]. Such discriminative approaches require a learning step, and the performance of the learned model generally depends on the size of the training dataset. Therefore, efficient learning algorithms are essential to handle the large-scale training dataset.

The space-time approach can be combined with GM to model the dependencies between neighboring gesture labels. In [18], the unary potential of the CRF model is defined by the simple linear function of the 2D silhouette or the 3D joint angle features. The overall model that includes the pairwise potential is then learned by optimizing the maximum likelihood objective function. Our method has two main differences from that of [18]. First, the unary term of our model is nonparametrically defined. Second, the SSVM framework rather than the maximum likelihood approach is used to learn the proposed model.

Several nonparametric space-time approaches are considered relevant to our method. In [31], a nonparametric action recognition method that is based on skeletal joint information is proposed, the EigenJoints descriptor that combines the posture, motion, and offset information is developed, and the Naive-Bayes-Nearest-Neighbor classifier is adopted to solve the gesture classification problem. The Moving Pose descriptor that captures skeletal joint position features and differential properties, such as the speed and acceleration of the joints, is proposed in [32]. The discriminative key frames for each action class are learned from the training dataset and are then used to produce the matching scores between the test sequence and the action classes. The Moving Pose descriptor is also applied to all frames according to the sliding window strategy for action detection in unsegmented sequences. Our approach can also handle unsegmented input video and produce temporally coherent frame-wise recognition results by global optimization with the smoothness constraint.

Gestures are elementary movements of a person's body part with the intend of representing meaningful information, and they constitute an interesting subspace of human activities [1], [33]. The *latest gesture recognition methods* are now reviewed as follows. In [34], a discriminative ferns ensemble approach is proposed for hand posture (i.e., static gesture) recognition from infra-red and depth images. The ferns ensemble model that has high capacity and minimal computation is trained discriminatively by minimizing the regularized hinge loss. It shows state-of-the-art accuracy and faster performance than other posture classification methods. However, it needs to be investigated whether the method can also be applied to gesture recognition from video. In [35], a domain-adaptive discriminative one-shot learning method is proposed for both localizing the gesture and classifying it into one of multiple classes. It is shown that a gesture classifier from a single example can be significantly enhanced by using a number of unlabeled training samples based on semi-supervised learning and domain adaptation. In [36], the gesture classification problem is addressed in the context of personalization, which involves learning the general classifier and adapting it for the intended user. For that purpose, a framework that learns a set of classifiers and selects best one is proposed and it shows more efficient and better performances than existing personalization methods. Our approach is based on the supervised learning with large-scale labeled data, therefore the problem setting of ours is quite different from those in [35] and [36].

We also investigate the top ranked methods in the *2014 ChaLearn LAP Challenge gesture recognition track*. The first-placed method [37] is based on a deep learning architecture with three data modalities, namely, depth information, grayscale video, and skeleton stream. The skeletal joint angles and normalized distances between joints are computed from the skeleton data as the skeleton based feature. For the depth and grayscale video, feature representation is automatically learned by the convolutional neural layers. The features are fused at the output layer of the neural architecture, and the results from the multiple temporal scales are then combined to produce the final prediction results. The second-placed method [38] extracts multi-modal features from skeleton, color, and depth data at multiple temporal scales to generate skeleton joint positions, rotations, velocities, and the HOG descriptors of the hands. The boosted classifiers are trained by using labeled training data and are then used to detect gestures in a one-versus-all sliding-window manner. The third-placed method [39] is the preliminary version of the proposed method in this paper. The previous approach is based on the CRF structure of the proposed model, but its parameter estimation depends on the brute-force search in the parameter space. In this paper, the CRF model is learned by using the efficient SSVM framework, which enables the use of more general and enriched models that cannot be learned in the previous method. Our extended model with structured learning achieves the state-of-the-art results for the gesture recognition track of the 2014 ChaLearn LAP Challenge.

3 Proposed Model

Suppose we are given a training dataset and each frame in a training sequence is labeled a gesture category $y^{(t)} \in \mathcal{Y} = \{1, \ldots, C\}$, where t and $\mathcal{Y}$ 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 with only one gesture category. The objective is to solve the *labeling problem* where each frame of a test sequence must be assigned a gesture category label.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

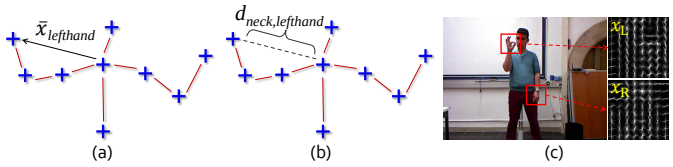


Fig. 2. Our feature extraction processes are illustrated. All skeletal joints are normalized with respect to the *neck* joint as shown in (a). They are concatenated to produce the *skeletal joint position feature*. As a viewpoint invariant feature, the Euclidean distances for all pairs of the skeletal joints are computed as in (b) to form the *skeletal joint distance feature*. To consider the detailed appearance of the left and right hands, HOG descriptors are extracted for the RGB image as in (c). By concatenating the HOG descriptors within several frames, the *appearance features* of the left and right hands are obtained.

3.1 Generative Probabilistic Model

Let us focus on classifying a single frame of a test multi-modal video. We assume that the gesture class at a particular frame depends not only on the observation of that frame, but also on a number of observations at its nearby frames. These long-range observations are used as the holistic feature for gesture classification. The generative probabilistic model for gesture classification can then be simply formalized as follows. The hidden random variables Y generate M multi-modal features $X_m, m = 1, \ldots, M$, where m denotes the type of the multi-modal features. The feature X_m is computed by considering the observations from 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 as follows:

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

We now 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. The skeletal joint features can be efficiently and robustly estimated from the depth image [4], [5]. Let $x_i, i = 1, \ldots, K$ denote the 3D coordinates of such joints. We then define the normalized joint coordinates $\bar{x}_i, i = 1, \ldots, K$ by taking the differences between x_i and the reference joint x_p, which is assumed to be the *neck* joint. To increase discriminability, we concatenate the normalized joint coordinates from the W_{P} frames near the current frame and construct the *skeletal joint position feature* $\mathbf{x}_{\mathrm{P}}$ as follows:

$$\mathbf{x}_{\mathrm{P}}^{(t)} = \left(\bar{x}_1^{(t-\lceil \frac{W_{\mathrm{P}}}{2} \rceil + 1)}, \ldots, \bar{x}_K^{(t+\lfloor \frac{W_{\mathrm{P}}}{2} \rfloor)} \right)^T. \tag{2}$$

The resultant $\mathbf{x}_{\mathrm{P}}$ is a $(W_{\mathrm{P}} \cdot 3 \cdot K)$-dimensional vector that holistically describes the motion dynamics of the human 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 $d_{i,j} = \|x_i - x_j\|$ between joints i and j. The *skeletal joint distance feature* $\mathbf{x}_{\mathrm{D}}$ is then defined as follows by concatenating all such distances for W_{D} frames:

$$\mathbf{x}_{\mathrm{D}}^{(t)} = \left(d_{1,2}^{(t-\lceil \frac{W_{\mathrm{D}}}{2} \rceil + 1)}, \ldots, d_{(K-1),K}^{(t+\lfloor \frac{W_{\mathrm{D}}}{2} \rfloor)} \right)^T. \tag{3}$$

Note that the dimensionality of $\mathbf{x}_{\mathrm{D}}$ is $W_{\mathrm{D}} \cdot \frac{K(K-1)}{2}$. $\mathbf{x}_{\mathrm{D}}$ is a relational pose feature [40] that describes the geometric relations between specific joints in a short sequence of frames.

We also consider the RGB image to exploit the details that are not captured by the skeletal joint features. The 3D joints of left and right hands are first projected to the RGB image. HOG descriptors x_{L} and x_{R} are then computed for the windows that are centered on the projected points. We concatenate the HOG descriptors of the W_{A} frames near the current frame as follows to construct our *appearance feature* $\mathbf{x}_{\mathrm{L}}$ for the left hand:

$$\mathbf{x}_{\mathrm{L}}^{(t)} = \left(x_L^{(t-\lceil \frac{W_{\mathrm{A}}}{2} \rceil + 1)}, \ldots, x_L^{(t+\lfloor \frac{W_{\mathrm{A}}}{2} \rfloor)} \right)^T. \tag{4}$$

The appearance feature for the right hand $\mathbf{x}_{\mathrm{R}}$ is analogously defined from the HOG descriptors that correspond to the right hand.

Given that our features are constructed from multiple frames, their dimensionality is generally high, especially for the HOG-based appearance features. Therefore, we use the principal component analysis (PCA) to reduce the computational complexity of our method. We also apply the standardization process to compensate for the different scales of the multi-modal features. As a result, each multi-modal feature will have zero mean and unit variance. Our feature extraction processes are illustrated in Fig. 2.

3.3 Active Hand Approach

According to the linguistics literature [41], there are two conditions on the formation of American Sign Language (ASL). The first is the *symmetry condition*, which states that if both hands move, they will have the same handshape and type of movement. The second is the *dominance condition*, and it states that if each hand has a different handshape, only the active or dominant hand can move. Inspired by this work, we aim to select the main hand and to use its appearance feature for gesture representation. We first introduce a new deterministic variable $a^{(t)}$ for each frame t:

$$a^{(t)} = \begin{cases} 0, & \text{if } e_l^{(t)} > e_r^{(t)}; \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

The energy of the left hand $e_l^{(t)}$ is defined as follows:

$$e_l^{(t)} = \frac{1}{W} \sum_{i=t-\lceil \frac{W}{2} \rceil + 1}^{t+\lfloor \frac{W}{2} \rfloor} \|x_l^{(i+1)} - x_l^{(i)}\|, \tag{6}$$

where W is the size of the box filter and $x_l^{(i)}$ denotes the 3D joint coordinate of the left hand. The energy of the right hand $e_r^{(t)}$ is analogously defined. The variable $a^{(t)}$ can then be intuitively understood as an indicator of which hand is more active at t-th frame. We hypothesize that the active hand is the main hand and that the feature of this hand is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

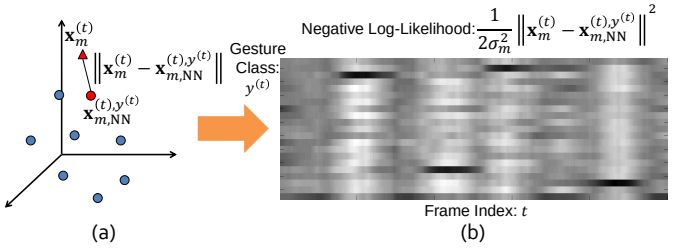


Fig. 3. This figure shows the likelihood estimation step in the proposed method. The training samples of the m-th multi-modal feature belonging to the gesture class $y^{(t)}$ are represented by circles in (a). They are matched to the input feature $\mathbf{x}_m^{(t)}$ (denoted by the red triangle) and the nearest neighbor vector $\mathbf{x}_{m,\mathrm{NN}}^{(t),y^{(t)}}$ (denoted by the red circle) is obtained. The Euclidean distance $\|\mathbf{x}_m^{(t)}-\mathbf{x}_{m,\mathrm{NN}}^{(t),y^{(t)}}\|$ is then computed and this process is repeated for all frame indexes $t=1,\ldots,T$ and gesture classes $y^{(t)}=1,\ldots,C$ to construct the negative log-likelihood matrix as in (b). We can get M negative log-likelihood matrices for all feature types $m=1,\ldots,M$ and then add all these matrices to produce the unary term in our CRF model as shown in the Fig. 1d.

helpful for gesture classification. We define the *appearance feature for the active hand* $\mathbf{x}_{\mathrm{A}}$ as follows:

$$\mathbf{x}_{\mathrm{A}}^{(t)}=\begin{cases}\mathbf{x}_{\mathrm{L}}^{(t)}, & \text{if } a^{(t)}=0;\\ \mathbf{x}_{\mathrm{R}}^{(t)}, & \text{if } a^{(t)}=1.\end{cases} \tag{7}$$

The active hand feature $\mathbf{x}_{\mathrm{A}}$ is adopted in this paper instead of the left and right hand features $\mathbf{x}_{\mathrm{L}}$ and $\mathbf{x}_{\mathrm{R}}$.

3.4 Nonparametric Estimation of Likelihood Functions

We now discuss how the likelihood function for each feature can be estimated from the training dataset. Let $\mathbf{x}_1^y,\ldots,\mathbf{x}_N^y$ denote all the features that are labeled a gesture category class y from all the training sequences. The *kernel density estimator* of the likelihood function is computed as follows:

$$\hat{p}(\mathbf{x}|y)=\frac{1}{N}\sum_{i=1}^{N}K(\mathbf{x}-\mathbf{x}_i^y), \tag{8}$$

where $K(\cdot)$ is the kernel function that must be non-negative and integrated to one. 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.

Given that N (i.e., the number of training samples belonging to each gesture class) is large, computing the likelihood in (8) is computationally expensive. Therefore, we approximate such likelihood by considering only the largest term in the summation (8). Given that the Gaussian kernel is assumed, this term corresponds to the nearest neighbor of the feature vector $\mathbf{x}$ within $\mathbf{x}_1^y,\ldots\mathbf{x}_N^y$, and the likelihood function (8) can be rewritten as follows:

$$\hat{p}^{\mathrm{NN}}(\mathbf{x}|y)\propto\exp(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}_{\mathrm{NN}}^y\|^2), \tag{9}$$

where $\mathbf{x}_{\mathrm{NN}}^y$ denotes the nearest neighbor vector. It can be efficiently found by using various approximate nearest neighbor search algorithms that include the randomized kd-trees [42]. We can consider the k-nearest neighbors ($k\geq 2$), but empirically, this improves the performance little.

We apply the above estimation process to all multi-modal features and obtain their corresponding approximate likelihoods. The combined likelihood function can then be computed and its negative log-likelihood can be written as follows:

$$L(\mathbf{x}|y)=\sum_{m=1}^{M}L_m(\mathbf{x}_m|y)=\sum_{m=1}^{M}\frac{1}{2\sigma_m^2}\|\mathbf{x}_m-\mathbf{x}_{m,\mathrm{NN}}^y\|^2, \tag{10}$$

where σ_m is the bandwidth for the m-th multi-modal feature and $\mathbf{x}_{m,\mathrm{NN}}^y$ denotes the nearest neighbor of the multi-modal feature $\mathbf{x}_m$ within the training samples of the gesture class y. Note that the bandwidth parameters $\sigma_m, m=1,\ldots,M$ control the relative importance between the multi-modal features. Fig. 3 shows the likelihood estimation process that is based on multiple feature matching.

3.5 CRF Model with Temporal Coherence

Let us assume that the test sequence is given and its multi-modal features are $\mathbf{x}^{(t)}=(\mathbf{x}_1^{(t)},\ldots,\mathbf{x}_M^{(t)})$, $t=1,\ldots,T$, where T is the length of the test sequence. We can then locally perform gesture recognition for the test sequence by using the negative log-likelihood in (10). Specifically, for each frame t of the test sequence, the optimal gesture class $y^{(t)*}$ can be found by minimizing the negative log-likelihood as follows:

$$y^{(t)*}=\arg\min_{y^{(t)}}L(\mathbf{x}^{(t)}|y^{(t)}). \tag{11}$$

Given that this locally optimized solution may lack temporal coherence, we formulate the following CRF model to enhance the temporal coherence of the solution:

$$E(\mathbf{y};\mathbf{x})=\sum_{t=1}^{T}U(y^{(t)};\mathbf{x}^{(t)})+\lambda\sum_{t=1}^{T-1}V(y^{(t)},y^{(t+1)}), \tag{12}$$

where $\mathbf{y}=(y^{(1)},\ldots,y^{(T)})$ denotes the gesture label sequence. The unary term is defined by using the negative log-likelihood ratio of each multi-modal feature as follows:

$$U(y^{(t)};\mathbf{x}^{(t)})=\sum_{m=1}^{M}(L_m(\mathbf{x}_m^{(t)}|y^{(t)})-\min_{y\in\mathcal{Y}\setminus y^{(t)}}L_m(\mathbf{x}_m^{(t)}|y)). \tag{13}$$

This slightly improves the performance rather than using the negative log-likelihood. The pairwise term is defined as the following simple smoothness constraint:

$$V(y^{(t)},y^{(t+1)})=\mathbb{1}(y^{(t)}\neq y^{(t+1)}), \tag{14}$$

where $\mathbb{1}(\cdot)$ is an indicator function. The parameter λ in (12) controls the strength of the smoothness constraint. The final gesture label sequence $\mathbf{y}$ can now be obtained by minimizing the energy function (12). Given that our model is the 1D pairwise CRF, its optimal solution can be efficiently computed via *dynamic programming*.

3.6 Extended Model

The proposed CRF model can be represented in a linear form. We define the joint feature functions $f_m(\mathbf{x},\mathbf{y}), 1\leq m\leq M$ as follows:

$$f_m(\mathbf{x},\mathbf{y})=\sum_{t=1}^{T}(\|\mathbf{x}_m^{(t)}-\mathbf{x}_{m,\mathrm{NN}}^{(t),y^{(t)}}\|^2-\min_{y\in\mathcal{Y}\setminus y^{(t)}}\|\mathbf{x}_m^{(t)}-\mathbf{x}_{m,\mathrm{NN}}^{(t),y}\|^2), \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

and we define $f_0(\mathbf{y})$ as follows:

$$f_0(\mathbf{y}) = \sum_{t=1}^{T-1} \mathbb{1}(y^{(t)} \neq y^{(t+1)}). \tag{16}$$

The CRF energy in (12) can then be rewritten as follows:

$$E_1(\mathbf{y}; \mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m f_m(\mathbf{x}, \mathbf{y}) + w_0 f_0(\mathbf{y}), \tag{17}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_M)^T$ denotes the $(M+1)$-dimensional parameter vector of the proposed CRF model. We refer to this model as the *gesture labeling conditional random field* (GLCRF-I). w_0 and $w_1, \ldots, w_M$ correspond to λ and $\frac{1}{2\sigma_1^2}, \ldots, \frac{1}{2\sigma_M^2}$, respectively. In this model, the single parameter w_m is in charge of all input and output pairs $(\mathbf{x}_m^{(t)}, y^{(t)}), t = 1, \ldots, T$ corresponding to the m-th multi-modal feature. This means that the relative importance between the multi-modal features remains constant for all gesture categories. However, which feature is useful for classification could vary with gesture classes. Therefore, we generalize and enrich the GLCRF-I model by setting different parameters $w_{m,c}$ for each feature type m and gesture class c. The joint feature functions $g_{m,c}(\mathbf{x}, \mathbf{y}), 1 \leq m \leq M, 1 \leq c \leq C$ of the extended model can be defined as follows:

$$g_{m,c}(\mathbf{x}, \mathbf{y}) = \sum_{t \in \mathcal{G}_c} (\|\mathbf{x}_m^{(t)} - \mathbf{x}_{m,\mathrm{NN}}^{(t),c}\|^2 - \min_{y \in \mathcal{Y} \setminus c} \|\mathbf{x}_m^{(t)} - \mathbf{x}_{m,\mathrm{NN}}^{(t),y}\|^2), \tag{18}$$

where $\mathcal{G}_c = \{i : y^{(i)} = c\}$ denotes the set of all frame indexes at which the output gesture label is c. The energy function of the extended model (GLCRF-II) can then be written as follows:

$$E_2(\mathbf{y}; \mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} \sum_{c=1}^{C} w_{m,c} g_{m,c}(\mathbf{x}, \mathbf{y}) + w_0 f_0(\mathbf{y}). \tag{19}$$

The dimensionality of the parameter vector $\mathbf{w}$ in this model is $M \cdot C + 1$.

4 Learning Method

4.1 Empirical Risk Minimization

The goal of our learning problem is to find the optimal hypothesis $h \in \mathcal{H}$ from the hypothesis space $\mathcal{H}$ given the training set $\mathcal{T} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where N is the number of training samples. Such hypothesis can be derived by minimizing the *empirical risk* $R(h)$ as follows:

$$R(h) = \frac{1}{N} \sum_{i=1}^{N} \Delta(h(\mathbf{x}_i), \mathbf{y}_i), \tag{20}$$

where $\Delta(\cdot, \cdot)$ is the loss function that measures how different the prediction $h(\mathbf{x}_i)$ is from the ground-truth output $\mathbf{y}_i$. Our hypothesis can be written as follows:

$$h(\mathbf{x}; \mathbf{w}) = \arg\max_{\mathbf{y}' \in \tilde{\mathcal{Y}}} \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y}'), \tag{21}$$

where $\mathbf{w}$ and $\Psi(\cdot, \cdot)$ denote the model parameter and the joint feature representation, respectively, and $\tilde{\mathcal{Y}} = \mathcal{Y} \times \cdots \times \mathcal{Y} = \mathcal{Y}^T$ denotes the structured output space. In the case of GLCRF-II in (19), the joint feature can be expressed as $\Psi(\mathbf{x}, \mathbf{y}) = -(f_0(\mathbf{y}), g_{1,1}(\mathbf{x}, \mathbf{y}), \ldots, g_{M,C}(\mathbf{x}, \mathbf{y}))^T$.

However, using the empirical risk (20) directly is problematic because our model relies on nonparametric feature matching with the training data. Assume that the hypothesis $h_{\mathcal{T}}(\mathbf{x}; \mathbf{w})$ is defined based on the training set $\mathcal{T}$. Matching one of the training samples with the training data will then cause zero loss $\Delta(h_{\mathcal{T}}(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i) = 0$ regardless of the model parameter $\mathbf{w}$. Therefore, we define the following *leave-one-out* empirical risk:

$$R_{\mathrm{loo}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \Delta(h_i(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i), \tag{22}$$

where the hypothesis $h_i(\mathbf{x}; \mathbf{w})$ is constructed by using the training data $\mathcal{T}_i = \mathcal{T} \setminus (\mathbf{x}_i, \mathbf{y}_i)$.

The loss function must be chosen before minimizing the leave-one-out empirical risk in (22). The negative conditional log-likelihood can be a candidate loss function as in [18] under the maximum likelihood principle. However, we find that minimizing such risk is intractable for the large-scale dataset in our experiments because the number of variables and labels is too large to compute the partition function. Therefore, we consider the Jaccard index (intersection-over-union), which is frequently used in semantic image segmentation [43], as an alternate standard loss function. This is also used as an evaluation measure for our experiments in Section 5. If the model is simple (i.e., the dimensionality of $\mathbf{w}$ is low), we can use simple brute-force search to find the optimal parameter. For example, in our GLCRF-I in (17) with the three features that are introduced in Sections 3.2 and 3.3, we can learn the model by searching the three-dimensional parameter space with suitable bound constraints. However, such process is infeasible for a highly sophisticated model with many parameters, such as our GLCRF-II in (19). In such cases, we incorporate the structured learning framework that is based on SSVM [44], [45].

4.2 Structured Learning Framework

In general, the empirical risk $R_{\mathrm{loo}}(\mathbf{w})$ in (22) is a complex non-convex function. Therefore, minimizing such function directly is difficult. The SSVM framework minimizes the surrogate function that is an upper bound of the loss function. For example, the loss function $\Delta(h_i(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i)$ is bounded from above by the *margin rescaling surrogate* $\sup_{\mathbf{y} \in \tilde{\mathcal{Y}}} \Delta(\mathbf{y}, \mathbf{y}_i) - \mathbf{w}^T \boldsymbol{\delta}\Psi_i(\mathbf{y})$, where we define $\boldsymbol{\delta}\Psi_i(\mathbf{y}) \equiv \Psi(\mathbf{x}_i, \mathbf{y}_i) - \Psi(\mathbf{x}_i, \mathbf{y})$. According to the margin rescaling formulation, the learning problem can be formulated as the following constrained quadratic program:

$$\begin{array}{l} \min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N} \sum_{i=1}^{N} \xi_i \\ \text{s.t. } \forall i, \forall \mathbf{y} \in \tilde{\mathcal{Y}} \setminus \mathbf{y}_i : \xi_i \geq \Delta(\mathbf{y}, \mathbf{y}_i) - \mathbf{w}^T \boldsymbol{\delta}\Psi_i(\mathbf{y}) \end{array}, \tag{23}$$

where C is a control parameter and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^T$ denotes the slack variables. However, the number of constraints in the above formulation is often huge because the cardinality of the structured output space $|\tilde{\mathcal{Y}}|$ may be extremely large. Therefore, the SSVM approach approximately solves the problem by using the cutting plane method [46]. In other words, the SSVM starts with an empty set of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

constraints and then incrementally constructs a lower approximation of the cost function by iteratively computing the most violated constraint (i.e., a cutting plane) and then adding such constraint to the constraint set. The learning algorithm is given in Algorithm 1.

Algorithm 1 Algorithm for learning the proposed model

Input: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, C, $\mathbf{w}_0$
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, N$
$\mathbf{w} \leftarrow \mathbf{w}_0$
repeat
  for $i = 1, \ldots, N$ **do**
    1. find the cutting plane $\bar{\mathbf{y}}_i$
    2. add $\bar{\mathbf{y}}_i$ to the constraint set:
      $S_i \leftarrow S_i \cup \{\bar{\mathbf{y}}_i\}$
    3. optimize with the new constraint set:
      $\min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^N \xi_i$
      s.t. $\forall i, \forall \mathbf{y} \in S_i : \xi_i \geq \Delta(\mathbf{y}, \mathbf{y}_i) - \mathbf{w}^T \boldsymbol{\delta}\Psi_i(\mathbf{y})$
  end for
until $\mathbf{w}$ is not changed (within a tolerance)

In the SSVM algorithm, the cutting plane can be obtained by solving the following *loss-augmented inference problem*:

$$\bar{\mathbf{y}}^i = \arg\max_{\mathbf{y}\in\tilde{\mathcal{Y}}} \Delta(\mathbf{y}, \mathbf{y}_i) - \mathbf{w}^T \boldsymbol{\delta}\Psi_i(\mathbf{y}). \tag{24}$$

This problem must be solved efficiently (in polynomial time), in which case the overall algorithm will have a polynomial runtime [45]. The difficulty of the problem depends on the structure of the loss function $\Delta(\mathbf{y}, \mathbf{y}_i)$ because in the case of our CRF model, the second term can be easily optimized via dynamic programming. Unfortunately, many useful loss functions, including the Jaccard index, involve high-order interactions, which result in an intractable high-order inference problem. Therefore, we use *Hamming loss* in this paper, which is expressed as follows:

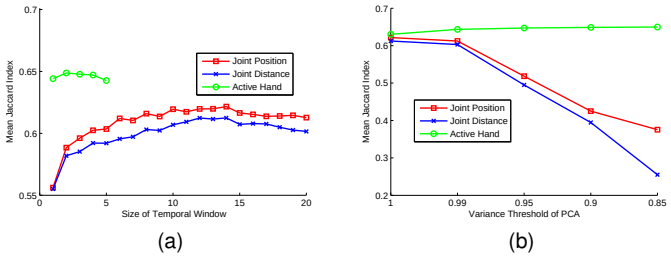
$$\Delta(\mathbf{y}, \mathbf{y}_i) = \sum_{t=1}^{T} \mathbb{1}(y^{(t)} \neq y_i^{(t)}), \tag{25}$$

where $\mathbf{y} = (y^{(1)}, \ldots, y^{(T)})$ and $\mathbf{y}_i = (y_i^{(1)}, \ldots, y_i^{(T)})$ are two gesture label sequences. The structure of this loss function is compatible with the joint feature representation of the proposed CRF model, specifically the expressions in (15) and (18). Therefore, the loss-augmented inference problem with the Hamming loss can be efficiently solved via dynamic programming.

5 Experimental Results

5.1 Dataset and Evaluation Metric

We use two different gesture datasets to evaluate the performance of the proposed gesture recognition method. The first dataset is the LAP dataset [12] that is introduced in ChaLearn LAP Challenge. It comprises 940 sequences (470 training, 230 validation, and 240 test sequences), with each sequence containing RGB, depth data, skeleton information that is extracted from the depth data by [4], and manually annotated gesture labels. The target gestures include 20 Italian cultural/anthropological signs. By containing a total of 13,858 gesture instances (7,754 training, 3,362 validation, and 2,742 test instances), this dataset has become among the largest-known datasets for gesture recognition. The second is the MSRC-12 Kinect gesture dataset[2] [47], which includes 594 skeletal joint sequences collected from 30 people performing 12 gestures. In total, there are 6,244 gesture instances. In the MSRC-12 dataset, ground-truth is provided in the form of an action point that is the characteristic point in time for the gesture rather than full frame labeling. We therefore assign the ground-truth gesture class to the frames within a fixed distance (15 frames) from the action point and use it as ground-truth labeling data.



(a) (b)

Fig. 4. The size of the temporal windows and the PCA variances for the features $\mathbf{x}_\mathrm{P}$, $\mathbf{x}_\mathrm{D}$, and $\mathbf{x}_\mathrm{A}$ are determined by simple greedy optimization with the training dataset. The resultant mean Jaccard index scores are plotted in (a) and (b).

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 the sets that include frames at which the n-th gesture is being performed in the s-th sequence. The Jaccard index can then be defined as follows:

$$J_{(s,n)} = \frac{|A_{(s,n)} \cap B_{(s,n)}|}{|A_{(s,n)} \cup B_{(s,n)}|}, \tag{26}$$

which represents the similarity between two sets. The Jaccard index $J_{(s,n)}$ is averaged over all gesture classes and sequences to produce the *mean Jaccard index*. We use this mean Jaccard index as the main evaluation criterion. We also compute the precision, recall, and F_1 score to evaluate the detection performance of our method. Therefore, we need to judge whether the detected gesture interval is the true/false positive. Similarly to object detection research [43], 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. The overlap ratio is computed as follows:

$$r = \frac{\mathrm{length}(I_\mathrm{gt} \cap I_\mathrm{p})}{\mathrm{length}(I_\mathrm{gt} \cup I_\mathrm{p})}, \tag{27}$$

where $I_\mathrm{gt} \cap I_\mathrm{p}$ represents the intersection of the ground-truth and predicted intervals, while $I_\mathrm{gt} \cup I_\mathrm{p}$ denotes their union. The precision, recall, and F_1 score are also averaged over all gesture classes to produce the *mean precision*, *mean recall*, and *mean F_1 score*, respectively.

5.2 Implementation Details

Each frame of the dataset sequence is labeled as one of the C gesture categories (20 for the LAP dataset and 12 for the

2. http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

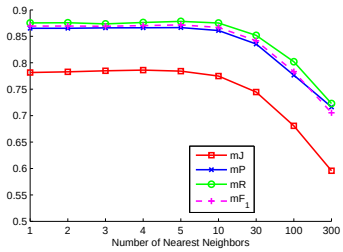


Fig. 5. Performance of the proposed method based on the skeletal joint position feature is illustrated with the number of nearest neighbors in the kNN approximation to the likelihood function. "mJ", "mP", "mR", "mF$_1$" denote the mean Jaccard index, mean precision, mean recall, and mean F$_1$ score, respectively. Note that the scale on the x-axis is not linear.

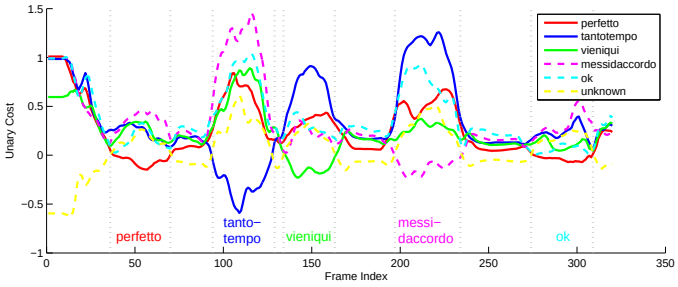


Fig. 6. Unary costs for one of the test sequences (sample index 701) are illustrated in this figure. Colored texts and gray dotted lines denote the ground-truth gesture classes and their starting/ending frames, respectively.

MSRC-12 dataset), but many frames in the dataset contain no meaningful gestures. We simply label these frames as an additional gesture category (i.e., *null*) in our gesture recognition framework. However, this additional gesture class is excluded from the computation of evaluation metrics, such as the Jaccard index, precision, recall, and F$_1$ score.

For the LAP dataset, $K = 10$ joints of the upper body (i.e., head, neck, L/R shoulder, L/R elbow, L/R wrist, L/R hand) are utilized for computing the skeletal joint based features $\mathbf{x}_{\mathrm{P}}$ and $\mathbf{x}_{\mathrm{D}}$. 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 $W = 10$ sized box filter is used to determine the active hand. The size of the temporal windows for constructing our features are $W_{\mathrm{P}} = 14$, $W_{\mathrm{D}} = 14$, and $W_{\mathrm{A}} = 2$. We then apply the PCA to each feature to reduce their dimensionality. The variance thresholds of the PCA are 0.99 for $\mathbf{x}_{\mathrm{P}}$ and $\mathbf{x}_{\mathrm{D}}$, and 0.85 for the appearance feature $\mathbf{x}_{\mathrm{A}}$ of the active hand. For determining these parameters, simple greedy optimization processes are conducted by using the training dataset. Fig. 4 shows the mean Jaccard index scores as a function of these parameters. For $\mathbf{x}_{\mathrm{P}}$ and $\mathbf{x}_{\mathrm{D}}$, the PCA variance 0.99 produces satisfactory results with the computational efficiency by the reduced dimensionality. Note that the aggressive PCA variance 0.85 results in better performance than other parameter settings in the case of $\mathbf{x}_{\mathrm{A}}$, which implies that $\mathbf{x}_{\mathrm{A}}$ is relatively more noisy than $\mathbf{x}_{\mathrm{P}}$ and $\mathbf{x}_{\mathrm{D}}$. Now, the remaining parameter is the control parameter C in the Algorithm 1, which is determined via cross validation. The open source VLFeat library [48] is used to compute the HOG descriptors and to perform fast nearest neighbor search with the randomized kd-trees. The SVM$^{\mathrm{struct}}$ framework [44] is used to conduct the structured learning of the proposed model.

TABLE 1
Performance of the proposed method based on a single feature is illustrated. "mJ", "mP", "mR", "mF$_1$" denote the mean Jaccard index, mean precision, mean recall, and mean F$_1$ score, respectively.

Model	Feature	mJ	mP	mR	mF$_1$
GLCRF-I	Joint Pos. (Local)	0.7547	0.7891	0.8556	0.8203
	Joint Position	0.7816	0.8653	0.8754	0.8694
	Joint Distance	0.7536	0.8403	0.8496	0.8443
	Left Hand	0.3613	0.6505	0.4421	0.5072
	Right Hand	0.6412	0.8270	0.7396	0.7785
	Both Hands	0.7136	0.9013	0.8494	0.8735
	Active Hand	0.7504	0.8885	0.8822	0.8845
BaseLinear	Joint Position	0.3073	0.4140	0.4080	0.3842
BaseRF	Joint Position	0.6277	0.7048	0.6953	0.6969

In the case of the MSRC-12 dataset, we utilize $K = 20$ joints of the full body (i.e., head, neck, L/R shoulder, L/R elbow, L/R wrist, L/R hand, spine, L/R/C hip, L/R knee, L/R ankle, L/R foot) to compute the skeletal joint position feature $\mathbf{x}_{\mathrm{P}}$. The skeletal joint distance feature $\mathbf{x}_{\mathrm{D}}$ is excluded because it does not improve the performance. The $W_{\mathrm{P}} = 56$ sized temporal window with a stride of 2 frames is exploited and the dimensionality of the computed feature is reduced by the PCA with the variance threshold 0.99.

5.3 Baseline Models

For comparative study, we have implemented two baseline models. They rely on the same features as our method and have similar CRF structures:

$$E_{\mathrm{baseline}}(\mathbf{y}; \mathbf{x}, \mathbf{w}) = \sum_{c=1}^{C} h_c(\mathbf{x}, \mathbf{y}) + w_0 f_0(\mathbf{y}), \qquad (28)$$

where the joint feature functions $h_c(\mathbf{x}, \mathbf{y}), 1 \leq c \leq C$ denote the unary term. The first model (BaseLinear) is based on the linear unary term, which is defined as the following linear function as in [18]:

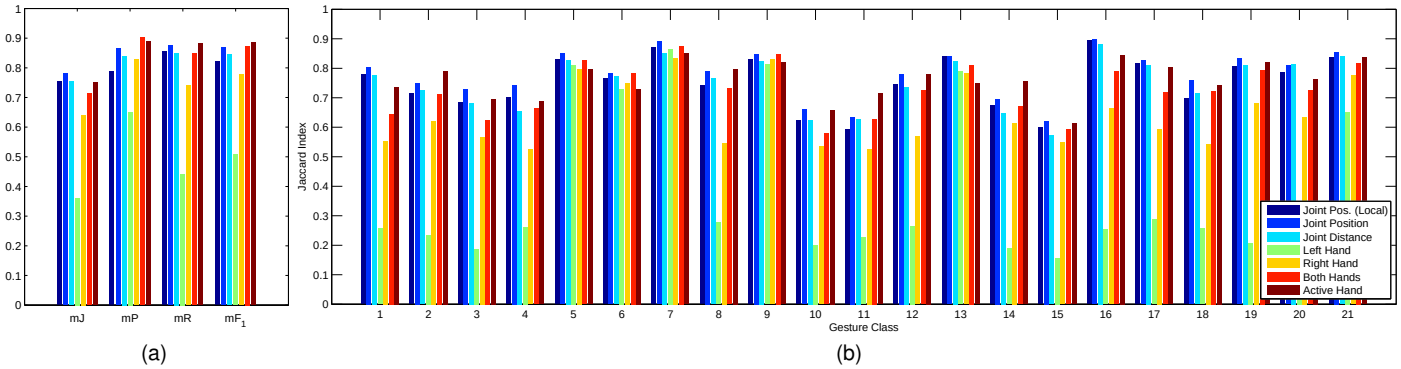
$$h_c(\mathbf{x}, \mathbf{y}) = \sum_{t \in \mathcal{G}_c} \mathbf{w}_c^T \tilde{\mathbf{x}}^{(t)}. \qquad (29)$$

In this equation, $\tilde{\mathbf{x}}^{(t)} = (\mathbf{x}^{(t)T}, 1)^T$ is the reparameterized feature vector and $\mathbf{w}_c$ is the linear weight corresponding to the gesture class c. For estimating the parameters $\mathbf{w}_c, 1 \leq c \leq C$ and w_0, we have used the structured learning framework with the Hamming loss. For the second model (BaseRF), the score of the random forest classifier is used as the nonlinear unary term:

$$h_c(\mathbf{x}, \mathbf{y}) = \sum_{t \in \mathcal{G}_c} p(y^{(t)} | \mathbf{x}^{(t)}), \qquad (30)$$

where $p(y^{(t)} | \mathbf{x}^{(t)})$ denotes the posterior class distribution computed by the random forest classifier. Similarly to [5], we employ simple comparison tests that compare the values of two coordinates of the feature vector, which produces better results than the decision stump test [47] in our experiments. The random forest classifier is trained by the standard greedy process [5], [47]. The hyperparameters (i.e., the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b)

Fig. 7. Performance of the proposed method based on a single feature is illustrated in (a). "mJ", "mP", "mR", "mF$_1$" denote the mean Jaccard index, mean precision, mean recall, and mean F$_1$ score, respectively. Jaccard index scores are illustrated for each gesture category in (b).

tree depth and the number of trees) and w_0 is determined by cross validation.

5.4 Performance Analysis on ChaLearn LAP Dataset

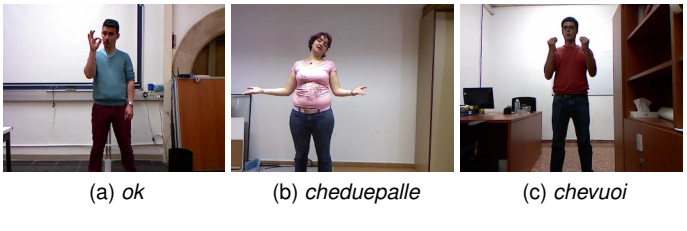
5.4.1 Single Feature

We now present the evaluation results of the proposed gesture recognition method. In this subsection, we consider the GLCRF-I in (17) with only a single feature. In this case, the model parameter $\mathbf{w} = (w_0, w_1)^T$ can be determined by a one-dimensional brute-force search. We first justify our nearest neighbor approximation (9) in the proposed model by conducting experiments using the skeletal joint position feature and the following k-nearest neighbor approximation to the likelihood function:

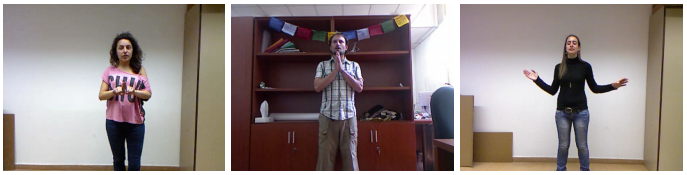
$$\hat{p}^{k\text{NN}}(\mathbf{x}|y) \propto \sum_{i=1}^{k} \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}^y_{\text{NN}_i}\|^2), \qquad (31)$$

where $\mathbf{x}^y_{\text{NN}_i}$ is the i-th nearest neighbor vector. Fig. 5 shows the results for the different values of k. The mean Jaccard index equals to 0.7862 in the best case ($k = 4$) and equals to 0.7816 when only the nearest neighbor ($k = 1$) is used. These results validate the effectiveness of our approximation.

To investigate the feasibility of our method, we compute the unary costs by using the skeletal joint position feature. Fig. 6 shows the unary costs for a test sample with the annotated ground-truth. The unary costs provide strong cues for gesture recognition. The value of the unary cost (i.e., negative log-likelihood) for the ground-truth gesture class is the lowest among all the gesture classes except the *ok* gesture in Fig. 6. Therefore, the proposed method can produce satisfactory results even when the coherence between neighboring frames is not considered. These findings are well illustrated in Table 1 and Fig. 7a, which show the evaluation results of our method with and without the smoothness term. The mean Jaccard index equals to 0.7547 according to the local approach (11). We then adopt the smoothness term in (14) and globally optimize the CRF energy in (12) to obtain a mean Jaccard index of 0.7816. The local approach achieves a reasonable performance, while the global approach demonstrates a further-improved performance.



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



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

Fig. 8. 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).

We then examine the performance of the proposed method according to the various features that are introduced in Sections 3.2 and 3.3. Table 1 and Fig. 7a show the evaluation results. The right-hand-based appearance feature demonstrates a better performance than the left-hand-based appearance feature because the right hand is more frequently used in many gesture instances of the dataset. Both hands are then simultaneously utilized, and this outperforms the single hand cases as expected. We then evaluate the active hand feature that is defined in Section 3.3, and the results of this feature are better than those that are obtained by using both hands. Unlike the use of both hands, only half the amount of information is required to represent the active hand feature.

For a more detailed analysis, the Jaccard index scores are computed for each gesture category. These scores are shown in Fig. 7b, where the numbers on the x-axis denote the following gesture classes: *vattene*, *vieniqui*, *perfetto*, *furbo*, *cheduepalle*, *chevuoi*, *daccordo*, *seipazzo*, *combinato*, *freganiente*, *ok*, *cosatifarei*, *basta*, *prendere*, *noncenepiu*, *fame*, *tantotempo*, *buonissimo*, *messidaccordo*, *sonostufo*, and *null*. The active-hand-based approach significantly outperforms the other approaches especially for the gesture category 11, i.e., *ok*, because the specific configuration of fingers is an important

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

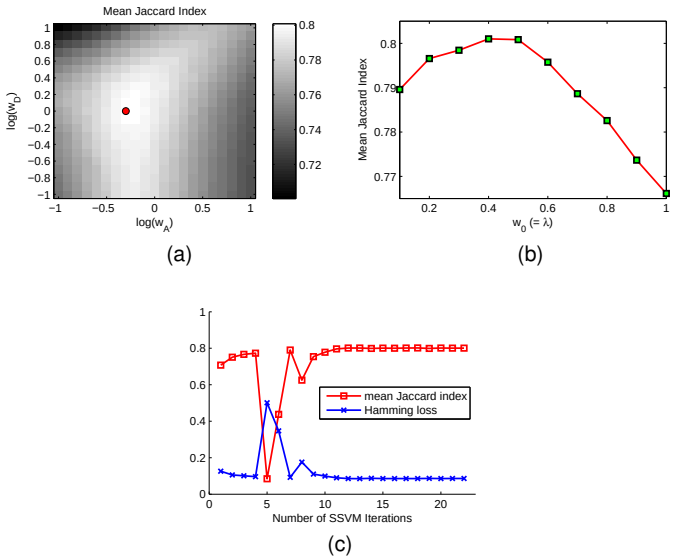


Fig. 9. The model parameters of our GLCRF-I are determined by the simple brute-force search in (a) and (b). It can also be done by using the SSVM algorithm as in (c).

characteristic that distinguishes *ok* from the other gestures. The left-hand-based method achieves comparable results with the right-hand-based method 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. Fig. 8 illustrates the abovementioned gestures.

5.4.2 Multiple Features

In this subsection, we investigate the performance of our multiple-feature-based approach. Let us consider the GLCRF-I with the joint position $\mathbf{x}_{\mathrm{P}}$, joint distance $\mathbf{x}_{\mathrm{D}}$, and active hand $\mathbf{x}_{\mathrm{A}}$ based features. The following parameters need to be determined in this case: w_{P}, w_{D}, w_{A}, and $w_0(=\lambda)$. Given the small number of model parameters, we can use the simple cross-validation process. Specifically, we set w_{P} to 1.0 and then optimize w_{D} and w_{A} via brute-force search. The smoothness parameter w_0 is analogously determined. Figs. 9a and 9b illustrate the results of the cross-validation process with the validation dataset. Table 2 shows that the mean Jaccard index score for the test dataset is slightly improved by optimizing the model parameters. Note that the simple feature combination with equal parameters (i.e., $w_{\mathrm{P}} = w_{\mathrm{D}} = w_{\mathrm{A}} = 1.0$) produces a Jaccard index of 0.8247, which is comparable with the optimal Jaccard index of 0.8268. This is thanks to the standardization process between the multiple features.

The GLCRF-I can also be learned by using the proposed learning method that is based on the SSVM framework. Fig. 9c shows the improvement in performance as the SSVM algorithm is converged. Given that the optimization function of our SSVM algorithm is the Hamming loss and not the mean Jaccard index, the converged mean Jaccard index score of this algorithm is not guaranteed to be optimal as compared to that of the cross-validation approach. Table 2 shows the final results for the test dataset. SSVM learning produces a slightly poor result (0.8239) than the cross-validation process (0.8268). In the case of the GLCRF-II with three features, the number of model parameters is $3 \times 21 + 1 = 64$, which makes the brute-force search infeasible. Therefore, we use the proposed structured learning method to find the optimal parameters, which generates a significantly improved result (0.8563) as shown in Table 2. The effectiveness of our extended model and learning approach is eventually validated.

Table 2 and Fig. 10 show the evaluation results when various combinations of multiple features are used. The use of multiple features significantly improves the recognition performance. The best mean Jaccard index result of 0.8563 is obtained when the joint position, joint distance, and active hand features are used altogether with the GLCRF-II model. This result has been ranked first place in the gesture recognition track of the ChaLearn LAP Challenge. The top-ranking results of the challenge are reported in Table 3. Note that the result of our simple nonparametric approach is comparable to that of the highly sophisticated method [37] based on deep learning that has proven to be the state-of-the-art feature learning technique in the various fields of computer vision. We believe this is because the skeletal joint based features themselves are already high-level representations suitable for gesture recognition.

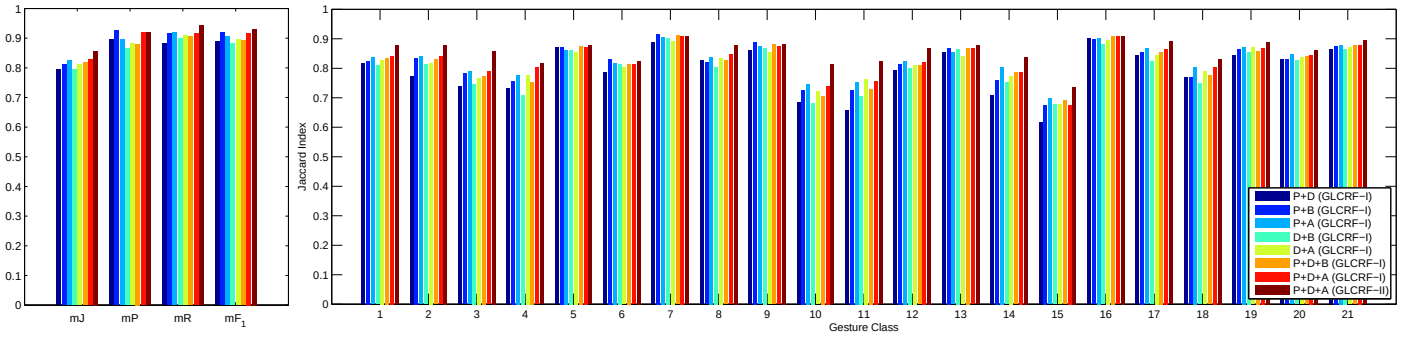
TABLE 2
Performance of the proposed method based on multiple features is illustrated. "P", "D", "B", and "A" denote the joint position, joint distance, both hands, and active hand based features, respectively. "None" denotes that learning is not performed and all model parameters are equally set to 1.0. "Holistic" means that all features are concatenated into a single vector for each frame and it is holistically used for gesture recognition.

Model	Feature	Learning	mJ	mP	mR	mF_1
GLCRF-I	P+D	CV	0.7948	0.8962	0.8824	0.8883
	P+B	CV	0.8110	0.9254	0.9157	0.9200
	P+A	CV	0.8244	0.8950	0.9191	0.9064
	D+B	CV	0.7952	0.8648	0.8996	0.8811
	D+A	CV	0.8114	0.8819	0.9087	0.8945
	P+D+B	CV	0.8182	0.8807	0.9056	0.8924
	P+D+A	CV	0.8268	0.9199	0.9158	0.9172
	P+D+A	None	0.8247	**0.9258**	0.9204	0.9226
	P+D+A	SSVM	0.8239	0.9145	0.9135	0.9134
GLCRF-II	P+D+A	SSVM	**0.8563**	0.9195	**0.9435**	**0.9310**
Holistic	P+D+A	CV	0.8170	0.9088	0.9121	0.9100

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

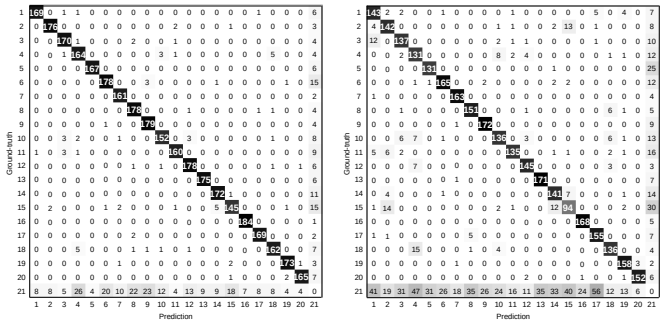
Rank	Team	Score
1	**GLCRF-II (SSVM)**	**0.8563**
2	Neverova et al. [37]	0.8500
3	Monnier et al. [38]	0.8339
4	**GLCRF-I (CV) [39]**	**0.8268**
5	Peng et al. [49]	0.7919
6	Pigou et al. [50]	0.7888
7	Wu [51]	0.7873
8	Camgoz et al. [52]	0.7466
9	Evangelidis et al. [53]	0.7454
10	Undisclosed authors	0.6888
11	Chen et al. [54]	0.6490

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b)

Fig. 10. Performance of the proposed method based on multiple features is illustrated in (a). "P", "D", "B", and "A" denote the joint position, joint distance, both hands, and active hand based features, respectively. Jaccard index scores are illustrated for each gesture category in (b).



(a) (b)

Fig. 11. The confusion matrices of the proposed GLCRF-II with the SSVM learning are illustrated. The original dataset is used in (a), whereas the newly constructed dataset for the cross-subject test is used in (b). Notice that the computation is performed based on per-interval evaluation.

The detection performance of the proposed method can be evaluated by comparing the predicted gesture intervals with the ground-truth intervals. The resultant precision, recall, and F_1 scores are reported in conjunction with the mean Jaccard index scores. The per-interval detection results show a similar tendency with the per-frame mean Jaccard index. The GLCRF-II model that is based on three features with the SSVM learning in Table 2 produces the best F_1 score (0.9310), which is defined as the harmonic mean of precision and recall. The corresponding confusion matrix is shown in Fig. 11a.

5.4.3 Comparison to Baseline Models

We have applied our two baseline models (i.e., BaseLinear and BaseRF) to the skeletal joint position feature $\mathbf{x}_{\mathrm{P}}$ of which the reduced dimensionality by PCA is $dim(\mathbf{x}_{\mathrm{P}}) = 42$. In the case of the BaseLinear model, the dimensionality of the model parameter $\mathbf{w} = (w_0, \mathbf{w}_1^T, \ldots \mathbf{w}_C^T)^T$ is $(dim(\mathbf{x}_{\mathrm{P}}) + 1) \cdot C + 1 = 904$, which is relatively high compared to our proposed GLCRF models. Training this baseline model takes approximately 5.97 hours, which is quite long due to the high dimensionality of the parameter vector. In Table 1, we can see the quantitative results of the baseline models, which are significantly outperformed by our methods. From the poor performance of the BaseLinear model, it is clear that the inherent nonlinearity of our problem cannot be handled by the simple linear unary term.

Both the proposed method and the baseline models share similar pairwise CRF structure that includes the simple smoothness constraint. The salient difference between them is in their unary costs. To show the effectiveness of the proposed nonparametric unary cost, we construct a frame-wise binary gesture detector for each class c by using the unary term $U(y^{(t)}; \mathbf{x}^{(t)})$ as follows:

$$S_c^{(t)} = \min_{y \in \mathcal{Y} \setminus c} U(y; \mathbf{x}^{(t)}) - U(c; \mathbf{x}^{(t)}), \tag{32}$$

where $S_c^{(t)}$ denotes a confidence score of the target gesture c. The per-class receiver operating characteristic (ROC) curves are plotted for 21 gesture classes in Fig. 12, which shows that our methods clearly outperform the baseline models.

5.4.4 Validity of Naive Bayes Assumption

Our proposed method relies on the naive Bayes assumption in (1). When the number of training samples is small, the naive Bayes assumption is known to increase the generalization capabilities of nonparametric classifiers [55]. To justify it for the large-scale LAP dataset, we have implemented another method where no conditional independence assumptions are enforced. In that model, our three features are concatenated into a single feature vector $\hat{\mathbf{x}} = (\mathbf{x}_{\mathrm{P}}^T, \mathbf{x}_{\mathrm{D}}^T, \mathbf{x}_{\mathrm{A}}^T)^T$, which is holistically used to estimate the negative log-likelihood by kernel density estimation as follows:

$$L_{\mathrm{holistic}}(\hat{\mathbf{x}}|y) = \|\hat{\mathbf{x}} - \hat{\mathbf{x}}_{\mathrm{NN}}^y\|^2, \tag{33}$$

where $\hat{\mathbf{x}}_{\mathrm{NN}}^y$ denotes the nearest neighbor of the combined feature $\hat{\mathbf{x}}$. As shown in Table 2, the CRF model composed of both $L_{\mathrm{holistic}}(\hat{\mathbf{x}}^{(t)}|y^{(t)})$ and the smoothness term in (14) produces slightly poor result (0.8170) compared to the GLCRF-I model. In the case of the GLCRF-II, the performance gap is relatively large and it is because the proposed SSVM based learning method can discriminatively learn the weights for feature types and gesture classes unlike the holistic feature based model.

5.4.5 Cross-Subject Test

The 940 sequences in the gesture dataset of the ChaLearn LAP Challenge are divided into training, validation, and test sets. These divisions are fixed and used throughout

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

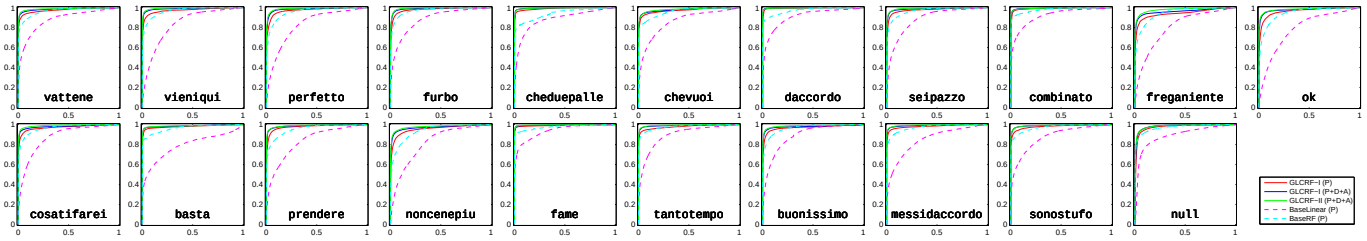


Fig. 12. The ROC curves of individual frame-wise binary classifiers are illustrated. The proposed method and the baseline models are evaluated for 21 gesture categories including the *null* class.

TABLE 4
Results of cross-subject test are illustrated.

Model	Feature	Learning	mJ	mP	mR	mF_1
GLCRF-I	P	None	0.5961	0.5445	0.6945	0.6081
	P+D+A	CV	0.6845	0.6890	0.7677	0.7240
GLCRF-II	P+D+A	SSVM	**0.7637**	**0.7942**	**0.8719**	**0.8299**
BaseLinear	P	SSVM	0.4005	0.5102	0.5233	0.4827
BaseRF	P	Greedy	0.4981	0.5461	0.5488	0.5370

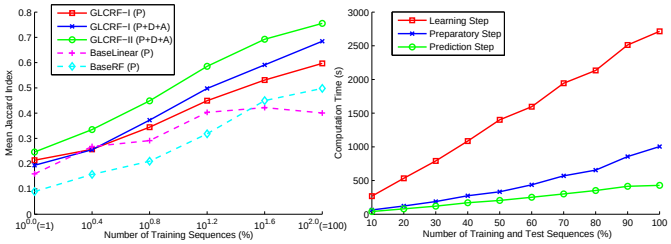


Fig. 13. The cross-subject generalization performance of the proposed method is plotted as a function of the different sizes of the training dataset in (a). Note that 246 test sequences are used for all cases. Our algorithm consists of the learning, preparatory, and prediction steps. The computation times of these steps are illustrated in (b), where 100% in the x-axis means that all 940 training and test sequences are used.

the challenge. In this original division, the same subject can appear in every set, hence making the dataset relatively easy to test. We therefore construct a dataset where a subject can only belong to one of the three subsets. Our new dataset comprises 468 training, 226 validation, and 246 test sequences. We then perform the cross-subject test by using this dataset to investigate the cross-subject generalization performance of the proposed method. The results of this test are reported in Table 4 and Fig. 11b. As expected, the performance of the model has decreased when the new dataset is used. The mean Jaccard index of GLCRF-II is decreased by 10.8 percent when the new dataset is used because such dataset contains a wide variety of body types and gesture styles for different subsets, hence resulting in more difficult and challenging dataset.

5.4.6 Size of Training Data

To investigate the learning power and the minimum training set size requirements for the proposed model, we report the performances of our method and the baseline models as a function of the different sizes of the training dataset in Fig. 13a. Our proposed method is based on the nonparametric approach by using a single nearest neighbor, therefore it could be vulnerable to overfitting. However the proposed method produces comparable or better results than the baseline models even when the size of the training data is small, which shows the robustness of our approach. We can see that the simple BaseLinear model has a high bias as the size of the training dataset grows. Whereas the performances of the proposed method and the BaseRF model continue to increase. This indicates that the two models have high capacity and there still remains room for improvement by adopting further training data.

5.4.7 Computational Complexity

We examine the computational complexity of the proposed method. A 3.00 GHz 8-core CPU machine with 64 GB RAM is utilized for that purpose. Our algorithm can be roughly divided into three parts, namely, (1) the parameter estimation step where the model parameters of the proposed CRFs are determined, (2) the preparatory step where the kd-trees are constructed from the training dataset, and (3) the prediction step that performs a nearest neighbor search with kd-trees and CRF optimization by dynamic programming. To investigate the scalability of our method, we conduct experiments by controlling the size of the training and test datasets. The runtime results for GLCRF-II with three features are shown in Fig. 13b, where the computing time increases linearly with the number of dataset sequences. Specifically, the results for three steps are 2714.9, 1026.0, and 431.5 seconds, respectively, when all 940 sequences are utilized. Note that the recognition of a test sequence takes approximately 1.4 milliseconds per frame to perform, hence showing the efficiency of our approach.

5.5 Performance Analysis on MSRC-12 Dataset

In this subsection, we present the evaluation results for the MSRC-12 dataset. For that purpose, we follow a *leave-persons-out* protocol, which means that a set of subjects is removed from the full dataset and is then employed as the test set. The proposed method is learned from the remaining training set and then the cross-subject generalization performance is evaluated by using the test set. Because a single feature (i.e., the skeletal joint position) is used for these experiments, the GLCRF-I model can be learned by a simple one-dimensional grid search for determining the smoothness parameter $w_0 (= \lambda)$ as in Fig. 9b. We first use a single subject for test and this process is repeated for all 30 subjects. The quantitative results are reported in Table 5. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2519021, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5
Evaluation results for MSRC-12 dataset are illustrated.

Model	Feature	Learning	mJ	mF_1
GLCRF-I	P	None	**0.6081±0.0724**	**0.6804±0.1081**
BaseLinear	P	SSVM	0.5595±0.0904	0.6770±0.1181
BaseRF	P	Greedy	0.5487±0.0852	0.6387±0.1068

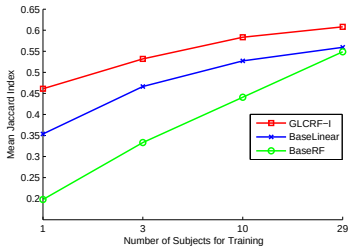


Fig. 14. The proposed method and the baseline models are evaluated by using the MSRC-12 dataset. The results are plotted as a function of the different numbers of subjects for training.

then change the number of training subjects and investigate the recognition performance for the remaining test subjects. Fig. 14 shows how the mean Jaccard index score changes with the varying training set size. For statistical tests, we have performed the paired t-tests using the mean Jaccard index results over 30 leave-one-person-out cross validation folds and verified that the improved performances with respect to the baseline models are significant at $\alpha = 0.05$.

The MSRC-12 dataset was originally introduced in [47], where the authors study different semiotic modalities of instructions including text, images, video, and combinations of them to investigate the most appropriate one for better performance of the learned gesture recognition system. The random forest classifier is adopted in [47] and the performance is evaluated by measuring the latency, that is, the time difference between the ground-truth action point and the predicted one. If it is smaller than a specified amount of tolerated latency δ, the prediction is regarded as a true positive. Given that the output of our method is provided in the form of time interval, we define the predicted temporal point as the middle of the interval. Quantitative results of two methods are illustrated in Table 6, which shows that the proposed method outperforms that of [47] for all five instruction modalities.

6 Conclusions

We have proposed a novel gesture recognition CRF model by using nonparametric multiple feature matching and developed its learning method according to the SSVM framework. Our approach can produce gesture category labels for all frames of the test sequence, which allows not only gesture classification, but also accurate localization. The experimental results demonstrate the convincing performance of the proposed method under various evaluation criteria. The target cost function of the proposed learning method is not the evaluation criterion itself (i.e., the mean Jaccard index), but the Hamming loss with a simple decomposable structure and can be easily optimized by using the SSVM algorithm. In the future, we aim to optimize directly the mean Jaccard index for improving the recognition performance. Our approach has been applied to recognize simple gestures that are composed of atomic motions with both hands. We aim to extend in the future our proposed framework for recognizing human actions with more complex and hierarchical structures.

TABLE 6
F_1 scores with tolerated latency ($\delta = 333ms$) are illustrated for five instruction modalities.

Instruction Modality	GLCRF-I	Fothergill et al. [47]
Text	0.591±0.128	0.479±0.104
Images	0.717±0.071	0.549±0.102
Video	0.746±0.049	0.627±0.053
Images+Text	0.703±0.084	0.563±0.045
Video+Text	0.769±0.074	0.679±0.035

Acknowledgments

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] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 16:1–16:43, 2011.
[2] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
[3] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer Berlin Heidelberg, 2013, pp. 149–187.
[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1297–1304.
[5] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.
[6] G. Johansson, "Visual motion perception," *Scientific American*, 1975.
[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2013, pp. 3192–3199.
[8] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.
[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
[11] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. European Conf. Computer Vision (ECCV)*, 2006, pp. 428–441.
[12] S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.
[13] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

[15] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[16] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011.

[17] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 329–349, 2013.

[18] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understaning*, vol. 104, pp. 210–220, 2006.

[19] C. Vogler and D. Metaxas, "Asl recognition based on a coupling between hmms and 3d motion analysis," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 1998, pp. 363–369.

[20] A. Jalal, M. Z. Uddin, J. T. Kim, and T.-S. Kim, "Recognition of human home activities via depth silhouettes and r transformation for smart homes," *Indoor and Built Environment*, pp. 1–7, 2011.

[21] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20–27.

[22] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1521–1527.

[23] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 842–849.

[24] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2009, pp. 444–451.

[25] F. Zhou and F. De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1282–1289.

[26] J. Wang and Y. Wu, "Learning maximum margin temporal warping for action recognition," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2013, pp. 2688–2695.

[27] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3265–3272.

[28] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[29] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 716–723.

[30] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu, "Concurrent action detection with structural prediction," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2013, pp. 3136–3143.

[31] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 14–19.

[32] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2013, pp. 2752–2759.

[33] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.

[34] E. Krupka, A. Vinnikov, B. Klein, A. Hillel, D. Freedman, and S. Stachniak, "Discriminative ferns ensemble for hand pose recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3670–3677.

[35] T. Pfister, J. Charles, and A. Zisserman, "Domain-adaptive discriminative one-shot learning of gestures," in *Proc. European Conf. Computer Vision (ECCV)*, 2014, pp. 814–829.

[36] A. Yao, L. Van Gool, and P. Kohli, "Gesture recognition portfolios for personalization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1923–1930.

[37] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[38] C. Monnier, S. German, and A. Ost, "A multi-scale boosted detector for efficient and robust gesture recognition," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[39] J. Y. Chang, "Nonparametric gesture labeling from multi-modal data," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[40] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 677–685, 2005.

[41] R. Battison, *Lexical Borrowing in American Sign Language*. Linstok Press, 1978.

[42] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application (VISAPP)*, 2009, pp. 331–340.

[43] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[44] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. International Conference on Machine Learning*, 2004.

[45] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[46] J. E. Kelley, "The cutting-plane method for solving convex programs," *Journal of the Society for Industrial & Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.

[47] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1737–1746.

[48] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[49] X. Peng, L. Wang, and Z. Cai, "Action and gesture temporal spotting with super vector representation," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[50] L. Pigou, S. Dieleman, and P.-J. Kindermans, "Sign language recognition using convolutional neural networks," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[51] D. Wu, "Deep dynamic neural networks for gesture segmentation and recognition," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[52] N. C. Camgoz, A. A. Kindiroglu, and L. Akarun, "Gesture recognition using template based random forest classifiers," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[53] G. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[54] G. Chen, D. Clarke, M. Giuliani, D. Weikersdorfer, and A. Knoll, "Multi-modality gesture detection and recognition with un-supervision, randomization and discrimination," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2014.

[55] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

Ju Yong Chang received the BS and PhD degrees in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea, in 2001 and 2008, respectively. From 2008 to 2009, he was a postdoctoral researcher at Mitsubishi Electric Research Lab. (MERL), Cambridge, MA. Currently, he is a senior researcher at Electronics and Telecommunications Research Institute (ETRI), Korea. His research interests include computer vision and machine learning, with a focus on human body pose estimation and human activity recognition.