Multi-modal Gesture Recognition Using Skeletal Joints and Motion Trail Model

Bin Liang(✉) and Lihong Zheng

Charles Sturt Universtiy, Wagga Wagga, Australia
{bliang,lzheng}@csu.edu.au

Abstract. This paper proposes a novel approach to multi-modal gesture recognition by using skeletal joints and motion trail model. The approach includes two modules, *i.e.* spotting and recognition. In the spotting module, a continuous gesture sequence is segmented into individual gesture intervals based on hand joint positions within a sliding window. In the recognition module, three models are combined to classify each gesture interval into one gesture category. For skeletal model, Hidden Markov Models (HMM) and Support Vector Machines (SVM) are adopted for classifying skeleton features. For depth maps and user masks, we employ 2D Motion Trail Model (2DMTM) for gesture representation to capture motion region information. SVM is then used to classify Pyramid Histograms of Oriented Gradient (PHOG) features from 2DMTM. These three models are complementary to each other. Finally, a fusion scheme incorporates the probability weights of each classifier for gesture recognition. The proposed approach is evaluated on the 2014 ChaLearn Multi-modal Gesture Recognition Challenge dataset. Experimental results demonstrate that the proposed approach using combined models outperforms single-modal approaches, and the recognition module can perform effectively on user-independent gesture recognition.

Keywords: Gesture recognition · Skeletal joints · HMM · SVM · 2DMTM · PHOG

1 Introduction

Human gesture recognition has been a very active research topic in the area of computer vision. It has been widely applied in a large variety of practical applications in real world, *e.g.* human-computer interaction, video surveillance, health-care and content-based video retrieval [9]. However, it is still a challenging problem owing to the large intra-class variability and inter-class similarity of gestures, cluttered background, motion blurring and illumination changes.

In the past decades, research on human gesture recognition mainly concentrates on recognizing human actions and gestures from video sequences captured by ordinary RGB cameras [1]. The difficulties of gesture recognition based on RGB video sequences come from several aspects. Human gestures captured by ordinary RGB cameras can only encode the information induced by the lateral

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 623–638, 2015.
DOI: 10.1007/978-3-319-16178-5_44

movement of the scene parallel to the image plane. Gesture motion information performed in a high dimensional space may be lost.

Recently, the launch of cost-effective depth cameras (*e.g.* Kinect) provides possibilities to alleviate the difficulties mentioned above. Depth information has long been regarded as an essential part of successful gesture recognition [11]. Using depth cameras, depth information can be obtained simultaneously with the RGB video. In addition, the positions of skeletal joints can also be predicted effectively from the depth data [22]. As a result, the depth maps and skeletal joints provide more information than RGB data. Thus, recent research has been motivated to explore more efficient multi-modal gesture recognition methods [2,5,17,28]. Furthermore, how to recognize human gestures using multi-modal information in an efficient way is still a hot topic.

In order to promote the research advance in gesture recognition, ChaLearn organized a challenge called "2014 Looking at People Challenge" [7] including three parallel challenge tracks. Track 3 (Gesture Recognition) is focused on multiple instances, user independent gesture spotting and learning. The dataset of the competition is recorded with a Microsoft Kinect camera, containing RGB videos, depth videos, user mask videos and skeleton data. Fig. 1 shows an example of different data sources available. The gesture vocabulary used in this dataset consists of 20 Italian cultural/anthropological signs. The most challenging points are that there are no obvious resting positions and the gestures are performed in continuous sequences. In addition, sequences may contain distracter gestures, which are not annotated since they are not included in the main vocabulary of 20 gestures [8].

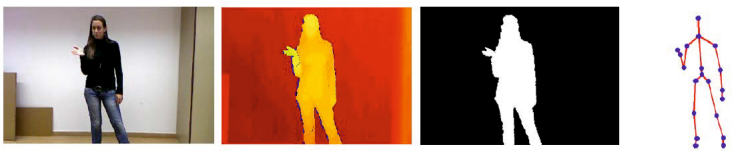


Fig. 1. An example from the dataset: RGB video, depth video, user mask video and skeleton data (left to right)

In this paper, we propose to use multi-modal data for gesture recognition. Specifically, a novel approach using skeletal joints and motion trail model [16] is proposed for multi-modal gesture recognition. The general framework of the proposed approach is illustrated in Fig. 2. In the gesture spotting module, within sliding windows we calculate the vertical difference of hand positions to divide one continuous gesture sequence into several gesture intervals. Three models, *i.e.* skeletal joints, depth maps and user masks, are then used to classify each gesture interval into one gesture category. For skeletal joints, *pairwise joints distance* and *bone orientation features* are extracted as skeleton features to encode 3D space information of the gesture. A concatenated classifier, HMM-SVM, is employed for skeleton features classification in time-domain. Furthermore, an RBF-SVM

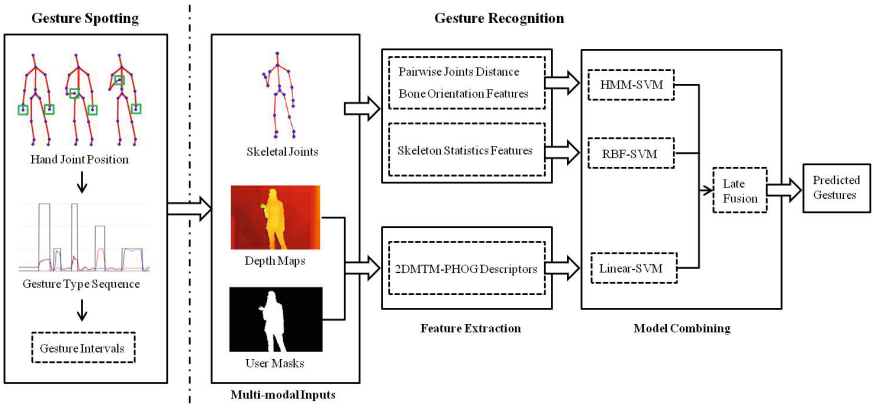


Fig. 2. The general framework of the proposed approach

classifier is used to classify *skeleton statistics features*. Meanwhile, depth maps and user masks are also used in our work, since skeleton data only encode joint information of human, ignoring the motion region information. Specifically, gesture regions are segmented by combining depth maps and user masks. 2D motion trail model (2DMTM) is performed on gesture regions for gesture representation, since 2DMTM is able to represent gesture motion information along with static posture information in 2D space to encode the motion region information of human gestures [16]. Then Pyramid Histograms of Oriented Gradient (PHOG) descriptors are extracted from 2DMTM. 2DMTM-PHOG descriptors are classified by linear-SVM. These models are complementary to each other. Finally, the probability scores from classifiers are fused for the final recognition. We evaluate our approach on 2014 Chalearn Multi-modal Gesture Recognition Challenge dataset [7] and further explain why the combination of the models achieves far better results than the single model.

The remainder of this paper is organized as follows. Section 2 reviews four categories of existing methods in the area of gesture recognition. In section 3, we provide a detailed procedure of our proposed approach based on skeletal joints and motion trail model. Experimental results and discussions are presented in section 4. At last, we give a conclusion of the paper and outline the future work in section 5.

2 Related Work

According to the data inputs, the existing methods of gesture recognition or action recognition can be roughly divided into four categories: *RGB video based*, *depth video based*, *skeleton data based* and *multi-modal data based*.

RGB video based methods. In video sequences captured by RGB cameras, the spatio-temporal interest points (STIPs) [13] are widely used in gesture recognition, as human gestures are showing spatio-temporal patterns. These methods

first detect interesting points and then extract features based on the detected local motion volumes. These features are then combined to model different gestures. In the literature, many spatio-temporal feature detectors [6,12,13,18] and features [14,21,26,27] have been proposed and shown promising performance for gesture recognition in RGB videos. Bobick and Davis [3] propose Motion History Image (MHI) and Motion Energy Image (MEI) to explicitly record shape changes for template matching. Tian *et al.* [23] employ Harris detector and local HOG descriptor on MHI to perform action recognition and detection. The core of these approaches is the detection and representation of spatio-temporal volumes.

Depth video based methods. With the release of depth cameras, there are many representative works for gesture recognition based on depth information. Li *et al.* [15] propose a bag of 3D points model for action recognition. A set of representative 3D points from the original depth data is sampled to characterize the action posture in each frame. The 3D points are then retrieved in depth maps according to the contour points. To address issues of noise and occlusions in the depth maps, Vieira *et al.* [24] present a novel feature descriptor, named Space-Time Occupancy Patterns (STOP). Yang *et al.* [31] develop Depth Motion Maps (DMM) to capture the aggregated temporal motion energies. The depth map is projected onto three pre-defined orthogonal Cartesian planes and then normalized. Oreifej and Liu [19] describe the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time (HON4D), depth, and spatial coordinates. Inspired by the great success of silhouette based methods developed for visual data, Jalal *et al.* [10] extract depth silhouettes to construct feature vectors. HMM is then utilized for recognition. More recently, Liang and Zheng [16] propose a three dimensional motion trail model (3D-MTM) to explicitly represent the dynamics and statics of gestures in 3D space using depth images. Specifically, depth images are projected onto two other planes to encode additional gesture information in 3D space. Evaluations on the MSR Action3D dataset [15] show a good performance.

Skeleton data based methods. Furthermore, motivated by the joints estimation of Kinect and associated SDK, there have been many different approaches relying on joint points for action recognition. In [25], the joints of the skeleton are used as interest points. In this way, the shapes of the area surrounding the joint along with the joint location information are captured using a local occupancy pattern feature and a pairwise distance feature, respectively. Xia *et al.* [29] propose a compact representation of postures named HOJ3D using the 3D skeletal joint locations from Kinect depth maps. They transfer skeleton joints into a spherical coordinate to achieve view-invariance. A more general method has been proposed by Yao *et al.* [32] where skeleton is encoded by relation pose features. These features describe geometric relations between specific joints in a single pose or a short sequence of poses. Yang *et al.* [30] propose a type of features by adopting the differences of joints. EigenJoints are then obtained by PCA for classification.

Multi-modal data based methods. Instead of using a single model, gesture recognition methods based on multi-modal data have been explored widely in recent

years. Zhu *et al.* [33] propose to recognize human actions based on a feature-level fusion of spatio-temporal features and skeleton joints. The random forest method is then applied to perform feature fusion, selection, and action classification together. Wu *et al.* [28], the winner team of the 2013 Multi-modal Gesture Recognition Challenge [8], propose a novel multi-modal continuous gesture recognition framework, which makes full exploration of both audio and skeleton data. A multi-modal gesture recognition system is developed in [17] for detecting as well as recognizing the gestures. The system adopts audio, RGB video, and skeleton joint models. Bayer and Silbermann [2] present an algorithm to recognize gestures by combining two data sources (audio and video) through weighted averaging, and demonstrate that the approach of combining information from two difference sources boosts the models performance significantly.

3 The Proposed Approach

We propose a multi-modal gesture recognition approach based on skeletal joints and motion trail model [16]. The framework consists of gesture spotting module and gesture recognition module. In gesture spotting module, each continuous gesture sequence is divided into gesture intervals using hand joint positions and sliding windows. After gesture spotting, classifiers based on skeleton features and 2DMTM-PHOG descriptors are constructed separately and then combined together to generate the final recognition result. In this section, we present the proposed approach in detail.

3.1 Gesture Spotting

The process of identifying the start and end points of continuous gesture sequences is called *Gesture Spotting*. We only focus on the positions of the hand joint to do gesture spotting, since all the gestures from the dataset are mainly performed using a single hand or two hands. The spotting method of work [17] is extended in our work by adding adaptive thresholds and sliding windows. The basic idea is that joint positions of two hands along vertical direction are varying while gesture is performing. Thus, the peaks of the hand joints position sequences indicate the presence of gesture performance, as shown in Fig. 3. In this way, a continuous gesture sequence can be segmented into individual gesture intervals according to the y-coordinates of hand joints. More specifically, the aim is to transform a continuous hand joint position sequence into a sequence of gesture types: left-hand dominant, right-hand dominant, two-hands dominant and neutral position. Therefore, the start and end points of gesture intervals would be the gesture boundaries in the gesture type sequence.

The gesture spotting in our approach consists of three steps. In the first step, single hand dominant gestures (left-hand dominant and right-hand dominant) are detected. A gesture type filter g is designed to transform hand joint position sequence to gesture type sequence. Let T be the total number of frames in a gesture sequence. Let $y(t)$ be the y-coordinate of a joint position in the t^{th}

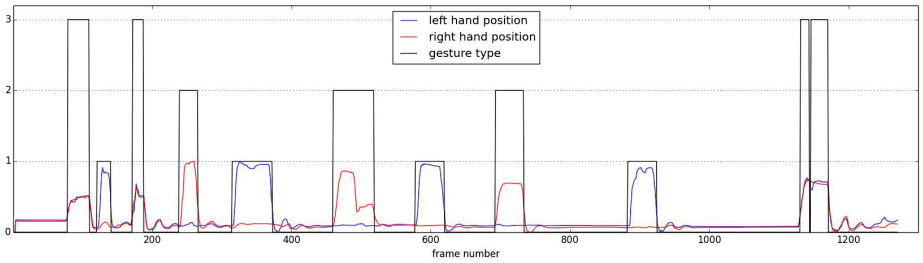


Fig. 3. Hand joint position and gesture type sequence

$(t = 1, 2, \ldots T)$ frame. In order to avoid the noise in the hand position sequence, a sliding window of size w is added here. Thus, a vector of a single joint position can be defined as $Y(t) = [y(t),\ y(t+1),\ \ldots,\ y(t+w-1)]$ to represent joint locations within a sliding window. The average joint position $\bar{Y}(t) = \|Y(t)\|_1/w$ can be obtained using L_1-norm. Therefore, the value of the filter g for the t^{th} frame is defined as follows:

$$g(t) = \begin{cases} 1, & \text{if } \bar{Y}_l(t) - \bar{Y}_r(t) > \eta_1 \\ 2, & \text{if } \bar{Y}_l(t) - \bar{Y}_r(t) < -\eta_1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\bar{Y}_l(t)$ and $\bar{Y}_r(t)$ are the average joint positions of left hand and right hand within the sliding window, respectively. $\eta_1 = d/5$ is an adaptive threshold, where d is the average distance between "shoulder center" and "hip center" within the sliding window. The filter value of 1 indicates left-hand dominant gesture, and the filter value of 2 indicates right-hand dominant gesture. If the value of the filter is 0, the gesture type could be either two-hands dominant or neutral position.

The second step is to identify the boundaries of two-hands dominant gestures. This is done by comparing the distance between hands position and hip position. If the distance is greater than the adaptive threshold $\eta_2 = d/10$, it is assumed that a two-hands dominant gesture is present. In this case, the filter value for that frame is changed to 3.

The last step of segmentation is to check the duration of the candidate gesture intervals. If the duration of a gesture is extremely short, the "gesture" will be discarded because noisy data could result in impulse intervals. On the other hand, if the duration is two times longer than the average duration, the "gesture" will be divided into two gestures. In general, this step is to discard "short-interval gestures" and separate "long-interval gestures". Fig. 3 shows the hand position sequences and the corresponding gesture type sequence after filtering. In the figure, hand joint positions are normalized to range [0, 1]. For gesture type sequence, the value of 1 indicates the left-hand dominant gesture, the value of 2 indicates the right-hand dominant gesture, the value of 3 indicates two-hands dominant gesture, and the value of 0 indicates neutral position.

3.2 Recognition based on Skeletal Joints

The spatial information of gestures is important for gesture recognition. Provided that the human skeleton joints can be estimated efficiently [22], we use pairwise joints distance and bone orientation features to complete the spatio-temporal features. The skeletal joints data provide 3D joint positions and bone orientations. Specifically, within frame t, skeleton data of each joint i is encoded using three coordinates: *world position* $s_{i,w}(t) = (x_{i,w}(t),\ y_{i,w}(t),\ z_{i,w}(t))$ representing the real-world position of a tracked joint, *pixel position* $s_{i,p}(t) = (x_{i,p}(t),\ y_{i,p}(t))$ with respect to the mapped pixel position over RGB maps, and *bone rotation* $s_{i,r}(t) = (w_{i,r}(t),\ x_{i,r}(t),\ y_{i,r}(t),\ z_{i,r}(t))$ in the form of a quaternion related to the bone rotation.

The skeleton positions of the joints are, however, not invariant to the gesture movements. Therefore, before extracting any features, all the 3D joint coordinates are transformed from the world coordinate system to a person centric coordinate system by placing the hip center at the origin. Inspired by the work of Wang *et al.* [25], we use pairwise joints distance as skeleton features. Furthermore, to eliminate the effect of variant sizes of the subjects, the transformed skeletons are then normalized by the sum of the pairwise distances of all selected skeletons. To characterize the posture information of the frame t, we compute the pairwise joints distances within that frame as follows:

$$f_w(t) = \left\{ \frac{\|s_{i,w}(t) - s_{j,w}(t)\|_2}{\sum \|s_{i,w}(t) - s_{j,w}(t)\|_2} \;\middle|\; i, j = 1, 2, \ldots N; i \neq j \right\} \tag{2}$$

where N is the number of selected joints. To capture the corresponding mapped posture information $f_p(t)$, the mapped pairwise joints distances within the frame t can be extracted in the similar way.

In addition, joint orientation information is essential for gesture movement. The orientation information is provided in form of quaternions. To encode the bone rotation features, we normalize the quaternion of joint i by its magnitude $\|s_{i,r}(t)\|_2$. According to our observation, most of the gestures are performed by upper-body movements, so we only extract 12 skeletal joints of upper-body from all 20 skeletal joints available: the head, shoulder center, spine, hip center, shoulders, elbows, wrists and the hands. Pairwise joints distance and bone orientation features are then concatenated together as the final skeleton features.

For classification task, we first use HMM [20] to construct a model for each gesture category. After forming the models for each category, we take a gesture interval and calculate its probability of all the 20 models. Then 20 probability scores are obtained. Usually the gesture is classified as one category which has the highest probability score, but we continue to use the obtained 20 scores as a new feature vector for the gesture, and adopt SVM [4] as a concatenated HMM-SVM classifier. The performance on validation data demonstrate that the concatenated HMM-SVM classifier performs better than the single HMM classifier.

The skeleton features mentioned above are extracted within each frame, and the holistic information (*e.g.* repeat movements, high-frequency motion regions

and the range of gesture movements) of gestures may be lost. Therefore, in order to capture the holistic information of skeleton features, another 4 statistics (variance, mean, minimum and maximum) are used to aggregate the skeleton features of each frame over a gesture interval. Then an RBF-SVM classifier is trained to classify the skeleton statistics features. In this way, gestures can be disriminated from each other.

3.3 Recognition based on Motion Trail Model

To capture gesture regions information, the 2D motion trail model (2DMTM) [16] is adopted in our approach. It employs four templates along the front view, *i.e.* depth motion history image (D-MHI$_f$), average motion image (AMI$_f$), static posture history image (SHI$_f$), and average static posture image (ASI$_f$). 2DMTM is able to represent the motion information and static posture information of human gestures in a compact and discriminative way.

In order to alleviate the influence of noisy background, we need to segment gesture regions from the original depth maps. Since user mask videos are available for each gesture sample, we propose to segment the gesture regions using depth videos and mask videos. Additionally, median filter is applied to remove the noise in the videos. We assume a binary user mask at frame t is M_t, and the corresponding depth map is D_t. Then gesture regions can be segmented by aligning these two maps to find their intersection $R_t = M_t \cap D_t$, as shown in Fig. 4

(a) User mask M_t (b) Depth map D_t (c) Gesture regions

Fig. 4. Gesture regions segmentation

Then the 2DMTM is performed on the segmented gesture region sequence. The motion update function $\Psi_{\mathrm{M}}(x, y, t)$ and static posture update function $\Psi_{\mathrm{S}}(x, y, t)$ are defined to represent the regions of motion and static posture with gesture performing. They are called for every frame analyzed in the gesture interval:

$$\Psi_{\mathrm{M}}(x, y, t) = \begin{cases} 1 & \text{if } \ K_t > \varsigma_M, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\Psi_{\mathrm{S}}(x, y, t) = \begin{cases} 1 & \text{if } \ R_t - K_t > \varsigma_S, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where x, y represent pixel position and t is time. $R_t = (R_1, R_2, \ldots, R_T)$ is a segmented gesture region sequence, and $K_t = (K_1, K_2, K_3, \ldots, K_T)$ is a difference

image sequence indicating the absolute difference of depth value. In addition, these two update functions need thresholds ς_M and ς_S for motion and static information within consecutive frames.

Therefore, the depth motion history image (D-MHI) $\mathrm{H_M}(x, y, t)$ can be obtained by using motion update function $\Psi_\mathrm{M}(x, y, t)$:

$$\mathrm{H_M}(x, y, t) = \begin{cases} T & \text{if } \Psi_\mathrm{M}(x, y, t) = 1 \\ \mathrm{H_M}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{5}$$

where T is the total number of frames in the gesture sequence. Additionally, static posture history image (SHI) $\mathrm{H_S}(x, y, t)$ can be generated utilizing the static posture update function $\Psi_\mathrm{S}(x, y, t)$ to compensate for static regions over the whole action sequence, which can be obtained in the similar way as D-MHI:

$$\mathrm{H_S}(x, y, t) = \begin{cases} T & \text{if } \Psi_\mathrm{S}(x, y, t) = 1 \\ \mathrm{H_S}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{6}$$

In order to cover the information of repetitive movements and repetitive static postures over the whole gesture interval, average motion image AMI and average static posture image ASI are employed. The summation of all motion information $\Psi_\mathrm{M}(x, y, t)$ or static information $\Psi_\mathrm{S}(x, y, t)$ and normalization of the pixel values define the AMI and ASI:

$$\mathrm{A_M} = \frac{1}{T} \sum_{t=1}^{T} \Psi_\mathrm{M}(x, y, t) \ , \quad \mathrm{A_S} = \frac{1}{T} \sum_{t=1}^{T} \Psi_\mathrm{S}(x, y, t) \tag{7}$$

Fig. 5 shows the motion trial model (2DMTM) of one gesture example. D-MHI and SHI present more recent moving regions and static regions brighter, respectively. AMI and ASI capture the average motion regions and average static regions information. Therefore, the 2DMTM gesture representation is able to characterize the accumulated motion and static regions distribution, meanwhile significantly reduces considerable data of depth maps to just four 2D gray-scale images.

(a) D-MHI (b) SHI (c) AMI (d) ASI

Fig. 5. 2DMTM of one gesture example

For feature extraction from 2DMTM, we apply PHOG [16] on 2DMTM to characterize local shapes at different spatial scales for gesture recognition.

Specifically, the 2DMTM-PHOG descriptor is extracted from the calculation of gradients in a dense grid of the 2DMTM to encode human gesture representation. It is directly performed on the four templates from the 2DMTM, which requires no edge or interesting regions extraction. In 2DMTM-PHOG, each template is divided into small spatial grids in a pyramid way at different pyramid levels. Each gradient orientation is quantized into B bins. Gradients over all the pixels within a grid are accumulated to form a local B bins 1-D histogram. Therefore, each template from 2DMTM at level l is represented by a $B \times 2^l \times 2^l$ dimension vector. Since there are four templates in 2DMTM, we concatenate the four PHOG vectors as the 2DMTM-PHOG descriptor. The obtained feature vector, $V \in \mathbb{R}^d$ ($d = 4 \times B \times \sum_{l=1}^{L} (2^l \times 2^l)$), is the 2DMTM-PHOG descriptor of the 2DMTM. In our experiment, we choose $B = 9$ bins and $L = 3$ levels empirically. Finally, linear-SVM is adopted to classify 2DMTM-PHOG descriptors.

3.4 Combining Skeleton and Motion Trail Model

In the above, three classifiers have been constructed for recognition: the spatio-temporal skeleton features based concatenated HMM-SVM classifier, skeleton statistics features based RBF-SVM classifier, and 2DMTM-PHOG descriptor based linear-SVM classifier. We have used LIBSVM [4] for all of our SVM implementations. LIBSVM has implemented an extension to SVM to provide probability estimates in addition to the decision values. Thus, each classifier is able to predict a probability score for each gesture category, indicating the confidence of prediction.

Table 1. Weights used for combining classifiers

Features	Classifier	Weight
Pairwise Joints Distance+Bone Orientation Features	HMM-SVM	0.35
Skeleton Statistics Features	RBF-SVM	0.45
2DMTM-PHOG Descriptors	Linear-SVM	0.20

We examine the influence of each classifier on performance and conclude a late fusion scheme to combine three classifiers based on their probability weights. Therefore, a fusion scheme is used to combine the weighted probability scores from the classifiers for the final recognition (Table 1). The weights are obtained by the experiments performed on the validation data.

4 Experimental Results

The proposed approach has been evaluated on the 2014 Chalearn Multi-modal Gesture Recognition Challenge dataset [7]. We extensively compare the proposed multi-modal approach with the single-modal approaches using mean Jaccard Index. In order to investigate the recognition module performance of the

proposed approach, we further use truth start and end points to do gesture spotting, and then evaluate gesture recognition using the proposed approach in terms of accuracy.

4.1 Dataset and Experimental settings

The 2014 Chalearn Multi-modal Gesture Recognition Challenge dataset is focused on "multiple instances, user independent learning" of gestures. There are 20 Italian sign categories, *e.g.*, *vattene*, *vieniqui*, and *perfetto*. Several features make this dataset extremely challenging, including the continuous sequences, the presence of distracter gestures, the relatively large number of categories, the length of the gestures sequences, and the variety of users [8]. Therefore, the dataset provides several models to attack such a difficult task, including RGB, depth videos, user mask videos, and skeletal model.

The dataset is split into three parts: training data, validation data, and test data. We use all the gesture sequences from training data for learning our models. Validation data is used for parameters optimization. The size of sliding window is 5 frames, and the number of hidden states in HMM is 15 in our work. Besides, the optimal parameters of SVMs are obtained by 5-fold cross-validation. At last, the proposed approach is evaluated on test data.

4.2 Evaluation Metric

For each unlabeled video sequence, the gesture category, corresponding start and end points are predicted using the proposed approach. Recognition performance is evaluated using the Jaccard Index. Therefore, for each one of the 20 gesture categories labeled for each gesture sequence, the Jaccard Index is defined as follows:

$$J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}} \tag{8}$$

where $A_{s,n}$ is the ground truth of gesture n in sequence s, and $B_{s,n}$ is the prediction for the corresponding gesture in sequence s. The proposed approach is evaluated upon mean Jaccard Index among all the gesture categories for all sequences. The mean Jaccard Index not only indicates the performance of recognition, but also the performance of gesture boundaries identification. Thus, higher mean Jaccard Index means better performance of the approach.

4.3 Comparison of Single-modal and Multi-modal Performance

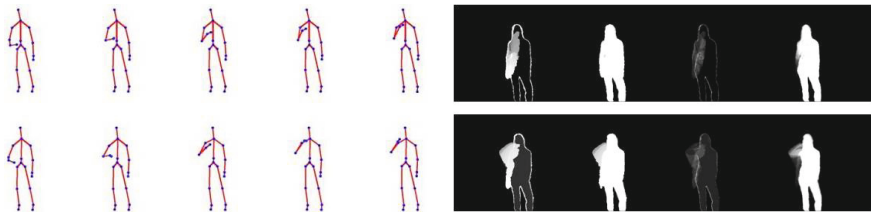
In order to evaluate the performance of the proposed approach, we compare the experimental results by using single-modal and multi-modal approaches. Continuous gesture sequences are first divided into individual gesture intervals using hand joints positions and sliding windows. Then we perform experiments using single-model and multi-modal approaches, and compare the results. For skeleton joints, pairwise joints distance and bone orientation features are classified

by concatenated HMM-SVM classifier, and skeleton statistics features are classified by an RBF-SVM classifier. Additionally, depth maps and user masks are used to segment gesture regions, and 2DMTM is employed for gesture representation. To capture the gesture regions information, 2DMTM-PHOG descriptors are adopted and a linear-SVM is trained for classification.

Table 2. Comparison of single-modal and multi-modal performance

Model	Classifier	Jaccard Index Score
Skeleton	HMM-SVM	0.453989
Skeleton	RBF-SVM	0.519499
Depth+Mask	Linear-SVM	0.462335
Skeleton+Depth+Mask	Multi-Modal	**0.597177**

The experimental results are shown in Table 2. From the results, we can see that the Jaccard Index scores are 0.453989, 0.519499 and 0.462335 using only HMM-SVM, RBF-SVM and linear-SVM, respectively. The RBF-SVM using skeleton statistics features has a higher score than other single-modal classifiers. It is probably because holistic skeleton information are user-independent, which means that the skeleton statistics features of the same gestures performed by different users are similar. In order to compensate the temporal information of skeleton joints, HMM-SVM is used to encode spatio-temporal skeleton features. Using our fusion scheme, the Jaccard Index score is improved to 0.597177, which is our final score for the competition.



(a) Skeleton model (b) Motion trial model

Fig. 6. Gesture samples with similar skeleton models

It is obvious that multi-modal recognition improves the performance significantly. Our analysis indicates that theses models are complementary to each other. Specifically, some skeleton joint positions are not very accurate in some cases. In addition, it is hard to differentiate two gestures when their skeleton models are very similar, but gesture representation using 2DMTM can provide discriminative information in this case. For example, as shown in Fig. 6(a), the two gesture samples have similar skeleton model, so the classifiers using skeleton model could fail to recognize them as two different categories. However, motion trail model provides complementary motion region information

of the gestures, which can help to complete recognition task in a better way (Fig. 6(b)). Therefore, the proposed multi-modal approach improves performance than single-modal approaches. According to the final ranking results released by the competition organizers, our team is ranked 11/17 in the final evaluation phase.

4.4 Recognition using Truth Spotting Labels

Gestures in the dataset are continuous and some of them are distracter gestures, so gesture spotting and distracter gestures rejection need to be considered in the competition. Thus, the performance of gesture spotting and distracter gestures rejection have a great effect on the final recognition results. In order to investigate how our approach impacts the final result, we first evaluate the performance of gesture spotting module using mean Jaccard Index, and the score is 0.819643. To investigate the performance of recognition module, we use the truth labels of start and end points provided by the dataset to do the spotting, and then recognize the gesture intervals using the multi-modal approach. In this way, only the performance of recognition module of multi-modal approach is evaluated, and compared with other single-modal approaches. Since we use truth labels to divide continuous gestures, the performance of recognition module is evaluated in terms of accuracy. The experimental results are shown in Table 3.

Table 3. Comparison of recognition using truth spotting labels

Model	Classifier	Accuracy
Skeleton	HMM-SVM	77.47%
Skeleton	RBF-SVM	83.02%
Depth+Mask	Linear-SVM	76.99%
Skeleton+Depth+Mask	Multi-Modal	**92.80%**

From the Table 3, we can see that the multi-modal approach also outperforms other single-modal approaches in recognition performance. The final recognition accuracy over the whole test data reaches **92.80%**, which is a relatively high score in "user independent" gesture recognition. Furthermore, the visualization of confusion matrix is illustrated in Fig. 7. From the confusion matrix, we can see that the highest accuracy is 100% for category 5 (*cheduepalle*), and the lowest accuracy is 85% for category 14 (*prendere*). Based on the results in section 4.3 and this section, we deduce that the spotting module remains to be improved to increase the overall performance of the proposed approach. In addition, we observe the test data and find out that the skeleton data of some gestures are missing, which causes some gestures are mis-discarded when performing gesture spotting. Therefore, only skeleton data might not be enough for a better gesture spotting, and combining other models could improve the performance.

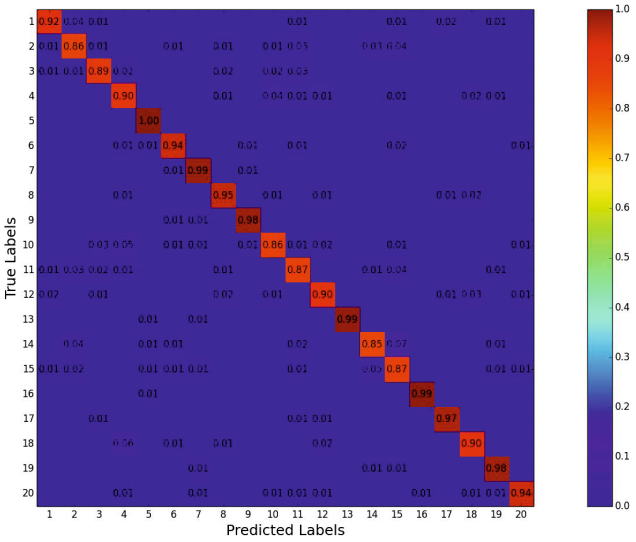


Fig. 7. Confusion matrix

5 Conclusion

We have presented a novel approach to multi-modal gesture recognition. The framework consists of gesture spotting module and gesture recognition module. In gesture spotting module, the start and end points of continuous gesture sequences are identified using hand joint positions within a sliding window. In gesture recognition module, the skeleton features characterize the spatio-temporal skeletal joints positions and holistic skeleton information while the 2DMTM-PHOG descriptors capture the motion regions information during a gesture performance. Three different classifiers, *i.e.* HMM-SVM, RBF-SVM and Linear-SVM, are then combined based on the late fusion scheme. We have conducted experiments on the 2014 Chalearn Multi-modal Gesture Recognition Challenge dataset, and shown that our proposed approach outperforms single-modal approaches. We further investigate the performance of recognition module in our approach, and get a relatively high accuracy of 92.80%. This demonstrates the good performance of recognition module. For the future work, we will combine more models to further improve the performance of gesture spotting under our proposed multi-modal framework.

References

1. Aggarwal, J., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) **43**(3), 16 (2011)
2. Bayer, I., Silbermann, T.: A multi modal approach to gesture recognition from audio and video data. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 461–466. ACM (2013)

3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3), 257–267 (2001)
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/cjlin/libsvm
5. Chen, X., Koskela, M.: Online rgb-d gesture recognition with extreme learning machines. In: Proceedings of the 15th ACM on International Conference On Multimodal Interaction, pp. 467–474. ACM (2013)
6. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2005, pp. 65–72. IEEE (2005)
7. Escalera, S., Bar, X., Gonzlez, 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: European Conference on Computer Vision Workshops (ECCVW) (2014)
8. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.: Multi-modal gesture recognition challenge 2013: Dataset and results. In: Proceedings of the 15th ACM on International Conference On Multimodal Interaction, pp. 445–452. ACM (2013)
9. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Computer Vision and Image Understanding **108**(1), 116–134 (2007)
10. Jalal, A., Uddin, M.Z., Kim, J.T., Kim, T.S.: Recognition of human home activities via depth silhouettes and transformation for smart homes. In: Indoor and Built Environment, p. 1420326X11423163 (2011)
11. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3d object dataset: Putting the kinect to work. In: Consumer Depth Cameras for Computer Vision, pp. 141–165. Springer (2013)
12. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
13. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2–3), 107–123 (2005)
14. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: MacLean, W.J. (ed.) SCVMA 2004. LNCS, vol. 3667, pp. 91–103. Springer, Heidelberg (2006)
15. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–14. IEEE (2010)
16. Liang, B., Zheng, L.: Three dimensional motion trail model for gesture recognition. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 684–691 (December 2013)
17. Nandakumar, K., Wan, K.W., Chan, S.M.A., Ng, W.Z.T., Wang, J.G., Yau, W.Y.: A multi-modal gesture recognition system using audio, video, and skeletal joint data. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 475–482. ACM (2013)
18. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **36**(3), 710–719 (2005)

19. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 716–723. IEEE (2013)
20. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
21. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, pp. 357–360. ACM (2007)
22. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1), 116–124 (2013)
23. Tian, Y., Cao, L., Liu, Z., Zhang, Z.: Hierarchical filtered motion for action recognition in crowded videos. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **42**(3), 313–323 (2012)
24. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.M.: STOP: space-time occupancy patterns for 3d action recognition from depth map sequences. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 252–259. Springer, Heidelberg (2012)
25. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297. IEEE (2012)
26. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
27. Wong, K.Y.K., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
28. Wu, J., Cheng, J., Zhao, C., Lu, H.: Fusing multi-modal features for gesture recognition. In: Proceedings of the 15th ACM on International Conference On Multimodal Interaction, pp. 453–460. ACM (2013)
29. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27. IEEE (2012)
30. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 14–19. IEEE (2012)
31. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM International Conference on Multimedia. pp. 1057–1060. ACM (2012)
32. Yao, A., Gall, J., Van Gool, L.: Coupled action recognition and pose estimation from multiple views. International Journal of Computer Vision **100**(1), 16–37 (2012)
33. Zhu, Y., Chen, W., Guo, G.: Fusing spatiotemporal features and joints for 3d action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 486–491 (June 2013)