Journal of Engineering Science and Technology
Vol. 17, No. 6 (2022) 4377 - 4397
© School of Engineering, Taylor's University

RANKED MULTI-VIEW SKELETAL VIDEO-BASED SIGN LANGUAGE RECOGNITION WITH TRIPLET LOSS EMBEDDINGS

SHAIK ASHRAF ALI, M. V. D. PRASAD, P. V. V. KISHORE*

Department of ECE, Koneru Lakshmiah Education Foundation, Vaddeswaram, AP, India
*Corresponding Author: pvvkishore@kluniversity.in

Abstract

Learning from multiview skeletal video data is difficult due to overlapping joints across views. In this work, we propose to overcome the above challenge by pairing views into positive intra and negative inter classes that are trained using a triplet loss embedding network. Further, the positive intra class views are ranked into two subgroups as view positive and support positive pairs through a view select network. Subsequently, the positive pairs are grouped with negative intra class views which are learned contrastively on a ranked multi-view deep metric learning network (RMVDML) using triplet and cross-entropy loss embeddings. This ensures highly discriminating class view features for classification on fully connected layers. Experimentations were conducted on 2D multi-view skeletal sign language videos and four benchmark action datasets. The proposed RMVDML has enhanced the efficiency of the skeletal video data for recognition tasks when compared to baselines.

Keywords: Deep metric learning, Sign language recognition, Skeletal video data, Triplet loss.

1. Introduction

The previous three decades have shown a tremendous improvement in the methods for visual recognition through the application of machine learning. One challenging use of the above area that has recorded a performance improvement was sign language recognition (SLR). The SLR was supported by computer vision and machine learning algorithms in different domains. Recent advances in the areas have boosted confidence in the SLR applications. Convolutional neural networks (CNNs) [1-3] and long short-term memory (LSTMs) [4, 5] are the two primary deep learning methods that elevated the confidence of the SLR systems. Though these deep models have shown higher recognition accuracies they could not assure a real-time deployable system. The main problem is with the hand and body movements during the signing process that will augment single view data into a multi-view data recovery problem. This kind of multi-view problems are commonly found in action recognition. However, multi-view problems can also occur in sign language due to subjects' movements or the camera positioning during capture.

Here we propose to expand the capacity of current deep metric learning (DML) [6] concepts to design a view-sensitive sign language recognition system. Generally, DML has succussed in the fields of speaker identification [7], face recognition [8], action recognition [9], satellite image classification [10], and person re-identification [11] problems. Here, DML is being investigated for solving multi-view recognition problems. The concept of DML used for SLR is unfolded in Fig. 1, where the model is trained to learn the within-class similarities and discriminate the across class dissimilarities. The learning process is instigated using a loss function defined by triplet loss. The triplet loss embedding is a distance metric that learns by maximizing the gap between within and between class features.

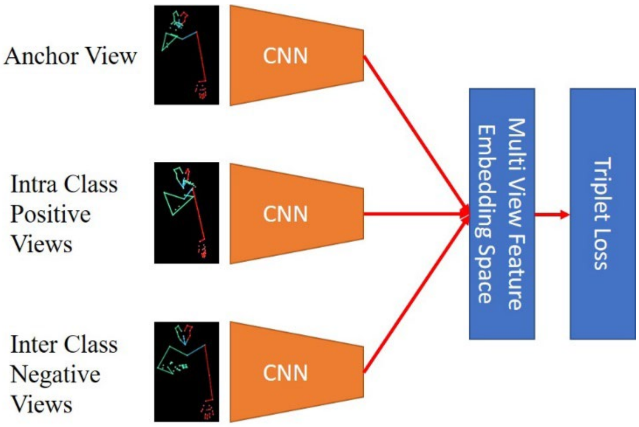
Anchor View
Intra Class Positive Views
Inter Class Negative Views
CNN
CNN
CNN
Multi View Feature Embedding Space
Triplet Loss

Fig. 1. Illustration of multi-view triplet loss embedding network architecture for sign language or action recognition.

Traditionally, DML pairs the entire labelled dataset as positive, negative and anchor pairs for training. As a result, each video sign or action class divides into video frames and transforming all these video frames into triplet pairs is a complex task. This process also increases the computational complexities of deep networks. Moreover, the numbering of triplet pairs in each class depends on the number of views available for processing. Compensating for the above problems we present a set based hard positive sample pairing mechanism which decreases the network loads for generating the triplet loss embedding for processing multi-view video data. For example, in a Nview sign language dataset, we have constructed an NC_2anchor - positive and anchor-negative pairs in each class. This learning process ranks the intra class views into two maximally dissimilar positive pairs. Subsequently, we construct NC_3 pairs for training the deep model with two positive intra class pairs and a set of inter class negative pairs. Significantly, the proposed model has reduced complexity without compromising the recognition accuracy.

Accordingly, we propose an inter class multi-view hard positive sample pairing mechanism using a view select network (P-VSN). First, we extract the features using a regular CNN on all available views within a class label. The extracted multi-view features are paired by finding the cosine distance between the views. Then we select (*N*-*k*) views that have maximum distance between views from within the class, where$k$gives ranked views in the view positive set. We call them maximally dissimilar view positive set (VP). Consequently, the remaining views are grouped into a view positive support set (VPS), which will be used during the training of DML. Accordingly, the view negative set (VN) is constituted by intra class views between different classes in the dataset. To classify using deep metric learning (DML), training pairs are constructed with (VP, VPS) and (VN, VPS). Since the construction of a positive set has a ranking effect, we named our process Ranked Multi-View Deep Metric Learning (RMVDML). The proposed network learns by computing the triplet and cross-entropy loss embeddings on pairs to (VP, VPS) and (VN, VPS). This process will improve the performance of the multi-view classifier by maximizing the inter-class distance and minimizing intra class distance between views. The separation is handled by a margin parameter between the two distances. Figure 2 shows the proposed ranked multi-view triplet loss embedding on skeletal video datasets.

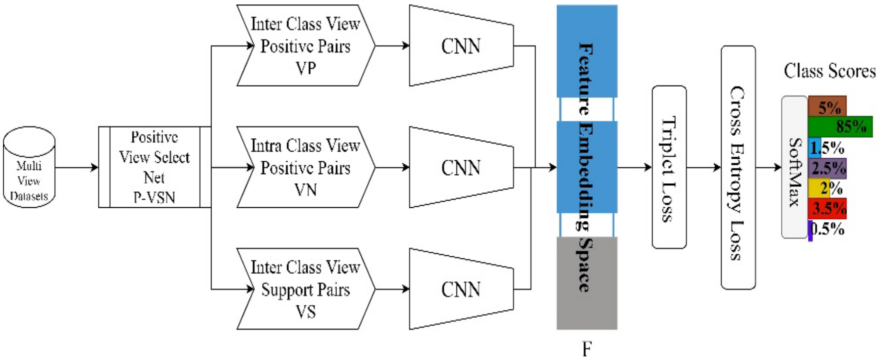
Multi View Datasets
Positive View Select Net P-VSN
Inter Class View Positive Pairs VP
Intra Class View Positive Pairs VN
Inter Class View Support Pairs VS
CNN
CNN
CNN
Feature Embedding Space
F
Triplet Loss
Cross Entropy Loss
SoftMax
Class Scores
5%
85%
1.5%
2.5%
2%
3.5%
0.5%

Fig. 2. Block representation ranked multi-view triplet loss embedding network architecture for skeletal video recognition tasks.

Finally, we test RMVDML on our own 200 class multi-view sign language dataset KLEF3DSL_2Dskeletal [12] with $N = 15$ views. Further, the designed ranked multi-view triplet loss embedding networks is being analysed on benchmark skeletal action datasets such as NTU RGB-D [13], SBU Kinect Interaction [14], KLYoga3D [15] and KL3D_MVaction [16]. Following this introduction is an overview of past methods with an insight into the advantages and disadvantages. The next section discusses methods applied for multi-view recognition of sign language. Finally, results and discussions were followed by conclusions that were drawn on the proposed solution for multi-view skeletal SL video recognition problems.

2. Literature Review

This section of the paper dwells on the advantages and disadvantages of the previous methods of sign language and action recognition in multiple views. Additionally, it also discusses the current models in deep metric learning.

SLR has been practiced in various forms based on data, features and classification algorithms [17]. The data usually comes from 3 sources, hand gloves (1D) [18], video cameras (2D) [19] and Kinect or leap motion (3D)[5]. The 4th and unique high-priced source are motion capture technology that has produced high precision synthetic sign language skeletal data [20]. Despite being a costly data source, the 3D motion captured signs exhibit naturalistic resemblance to real time human actions with far better representations than the other sources. However, the most commonly used source for research experimentations is 2D RGB video data [21]. A wide variety of algorithms were proposed in the last few decades for video pre-processing, feature extraction and recognition [22, 23]. Most of these algorithms solved some type of spatial, temporal or paired feature representation of video object data effectively. These features are further classified using all the traditional machine learning algorithms. The most popular classifiers were Hidden Markov Models(HMM) [24] and Artificial Neural Networks(ANN) [25].

With the advent of deep learning frameworks, the 2D video based SLR has become powerful with the option of feature learning rather than feature extraction. A large contingent of them are available for perusal [26]. The accuracies reported by these methods are not reproducible or they simply fail to generalize on the video quality or the signer. This has motivated researchers towards higher dimensional data such as RGB – D or 3D skeletal representations. Multimodal video sequences that are fed into multiple streams of Convolutional Neural Networks (CNNs) have been dominating this research field. Undoubtedly, the evidence points to exceptional performances in real-time for sign (action) recognition applications [27]. The recognition accuracies were better than the single modal datasets. However, the training requires higher computing powers, and the datasets are captured with special devices making it an unfeasible deployable solution.

Eventually, to develop a real-time SLR or HAR system, it is intuitive to learn multiple views across datasets. This has initiated action recognition research to move in the direction of developing view-based learning algorithms [28]. Multi-view HAR has evolved through research using dictionary learning [29,30], Neural Networks with adaptable views [31], CNN's [32] and deep attention models [33], to name a few. The most widely researched and acknowledged models are from deep learning networks. Moreover, visual attention models with deep CNNs have established themselves as a formidable solution to multi-view learning [34]. Despite their success, attention models are specific to a particular view and the view-specific features are to be fused accordingly for classification by the dense layers. The fusion mechanisms ensemble the view-specific features into a multi-view feature vector that has failed to capture view variations in multi-view data [35].

This motivated us to look for a more robust learning model that can learn to collaborate between views during training. Consequently, deep metric learning (DML) has shown to cluster highly similar within-class samples by learning the loss dynamics across different classes [36]. The loss dynamics is calculated using the contrastive [37] and triplet functions [38] for training. In the past few years, DML has become a driving force in multiple vision-based applications such as person identification, face recognition in the wild, speaker identification, image classification and remote sensing [36]. Additionally, there are multiple procedures in which the loss can be included in the objective function apart from triplet and contrastive techniques. Some of them are, hierarchical triplet loss, hard triplet loss [39], angular loss [40] and n – pair multiclass loss [41]. All these losses have distinctive advantages, especially in maximizing within class and minimizing across class similarities for utmost performance. Lastly, these losses are difficult to implement due to multiple regularizations that are specific to the application.

The objectives of the proposed work are threefold: 1. To extract the most relevant views from a large pool of views. 2. To compute triplet loss embedding across the paired views from inter and intra classes. 3. To classify multi-view skeletal signs using triplet and cross-entropy losses with a margin parameter. Contrastingly, the uniqueness of the proposed method when compared to the existing deep metric learning models is threefold: 1. Application of DML for multi-view skeletal video recognition with novel pairing mechanism. 2. The reduced complexity in triplet pairing for generating highly discriminative features is automated using the view select network (P-VSN). 3. Calculating set-based view distances in the proposed work as against sample-based distances in the existing triplet loss setting.

3. Methodology: Rank View Triplet Loss Embedding

This work aims to optimize the rank-based triplet loss embedding on multi-view skeletal sign(action) datasets by training deep networks for recognition. We first present the view select network (P-VSN) architecture followed by the procedure to create an optimized triplet pair for training the DML. Secondly, we derive the procedure to train a deep network on the derived training pairs using the triplet loss embedding and global cross-entropy loss functional.

3.1. View select network (P-VSN)

This section aims to develop the underlying theoretical background applied to select pairs for training triplet loss networks from a large pool of views. Let there be a set of N video views. The goal is to select a set of positive view pairs kC_2 $\forall\{k = \text{Ranked Pairs}\}$in pairs of two in ranked order. This is in contrast to the regular pairing mechanism where the number of positive pairs generated will beNC_2. The remaining views are then grouped into support set. We propose to automate the process of selecting high-ranked positive pairs using CNNs for feature extraction and consequently finding similarities in these feature spaces. The idea is to form positive pairs within a class that are dissimilar enough to act as an anchor and positive sample sets. Consequently, the support set forms the negative set with inter-class views. The triplet loss embeddings are computed on the positive pairs and the negative support samples to learn the view variance features that can help in decision making. The following Fig. 3 shows the architecture of view selective deep network (P-VSN) for generating automatic positive sample pairs across multiple views.

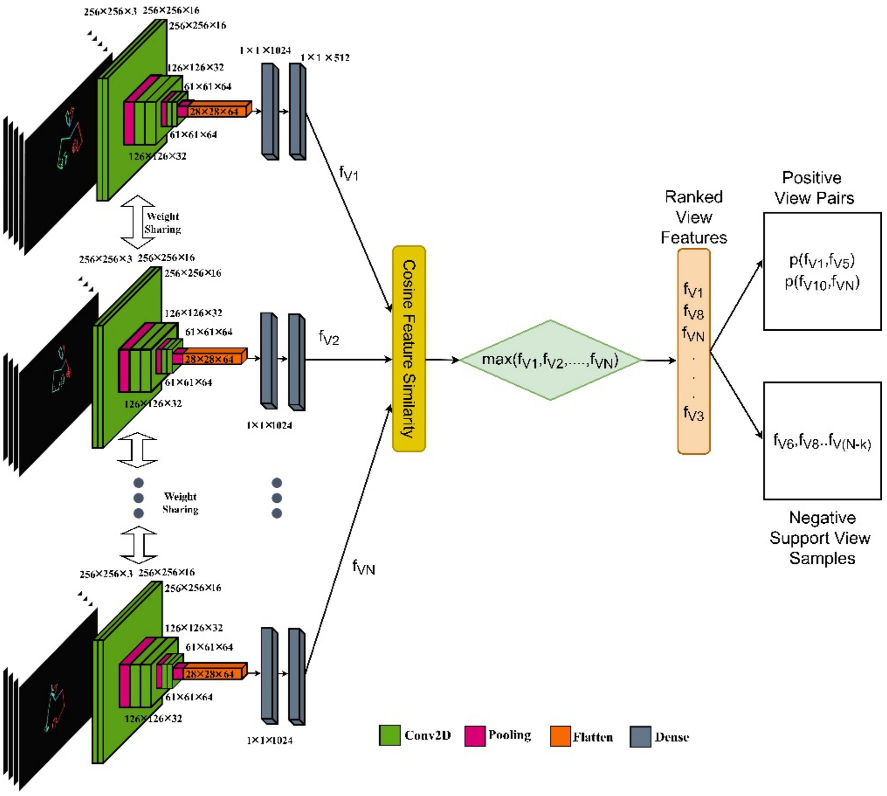

Fig. 3.The positive view select network architecture.

The architecture of CNN is incepted from trimmed VGG with 16 layers. However, our P-VSN has only 12 layers, i.e., 6 convolutional + ReLu, 3 Maximum pooling, 1 flatten and 2 dense layers. The strides across max pooling layers in kept constant at two, whereas it is one across convolutional layers. Contrasting to the regular multi-stream CNN models, we used a single stream CNN model that was accessed multiple times depending on the dataset views available for training. Let $x_v = \{V_v\}\forall v = 1\ to\ N$ be the RGB skeletal video sequences in$N$views with $V \in R^3$. The CNN model will extract the features from $x_v$with view specific labels y_v using the trainable parameters θ_{p-vsn} by optimizing a loss function L on the overall multi-view dataset as

$$\theta_{p-vsn} = arg \underset{\theta_{p-vsn}}{min}\ L\left(\theta_{p-vsn}; x_v, y_v\right) \qquad (1)$$

The trained model θ_{p-vsn} has view specific features f_v at the output of the dense layers as

$$\{f_v\}_{v=\{1,N\}} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_v(i,j) * K(k-i, k-j)\forall k \in \text{kernel size} \qquad (2)$$

The features from each of the views are extracted and cosine similarity is calculated on all possible view pairs. For a Nview data, we will have NC_2 feature pairs. The cosine similarity $C_{(v,v+1)}$ between view feature pairs with n attributes in each feature vector is computed as

$$\cos(f_v, f_{v+1}) = C_{(v,v+1)} = \frac{\sum_{i=1}^{n} f_v^i f_{v+1}^i}{\sqrt{\sum_{i=1}^{n} f_v^i}\sqrt{\sum_{i=1}^{n} f_{v+1}^i}} \tag{3}$$

The cosine similarity scores are ranked in chronological order as

$$C_f(p) = arg \min_{\forall i=1\ to\ NC_2} (C_v, C_{v+1})_i \tag{4}$$

where p points to the pair placeholder which contains the two views with maximum distance. Here, we select the top pair that has minimum similarity or maximum dissimilarity between them. These highly dissimilar within class pairs act as anchor positive pairs $\{F_a, F_p\}$. We can also use top P pairs for training. Specifically, the remaining samples will be grouped into a negative support set F_n. Subsequently, iterating the above process over the entire dataset results in inter class positive pairs and intra class support negative class. Even though there are few positive samples in support class, we will consider only intra class features during the training of each set. The complete view sample sets for training the deep metric learning model will be $\{F_a, F_p, F_n\}$. The reconstituted triplet pairs will be applied as input to the ranked multi-view deep metric learning network (RMVDML) for learning the multi-view features.

This process ensures that the positive anchor pairs within the class features are pulled closer while the anchor negative pairs across the classes are pushed farther by a margin parameter during training. The following subsection presents a bird's eye view of the designed architecture and the advantages it offers over the traditional multi-stream CNN models.

3.2. Ranked multi-view deep metric leaning (RMVDML)

This implementation aims to maximize the recognition accuracy by pulling the highly relatable within class views from discriminating them from uncorrelated inter-class views. The primary function of metric learning is similarity measurement between pairs of data samples by preserving the distance metrics. There are two types of metric learning models followed extensively: contrastive [36] and triplet loss [37].

Given a pair of features in embedding space $(f_{vi}, f_{vj})$from different classes in the dataset, the contrastive loss l_c is calculated with the cost function defined as

$$l_c(f_{vi}, f_{vj}, y_{ij}) = \sum_{i,j} y_{ij} d^2(f_{vi}, f_{vj}) + (1 - y_{ij}) h\left(\delta - d(f_{vi}, f_{vj})\right)^2 \tag{5}$$

where h is the hinge loss operator defined as

$$h(f_v) = max(f_v, 0) \tag{6}$$

The class label indicator y_{ij} on the trained model parameter θ is defined as

$$y_{ij} = \begin{cases} 1 \ \ \forall \ \ \left(\theta(f_{vi}) = \theta(f_{vj})\right) \\ 0 \ \ \forall \ \ \left(\theta(f_{vi}) \neq \theta(f_{vj})\right) \end{cases} \tag{7}$$

Finally, the parameter $d(f_{vi}, f_{vj})$ is the Euclidian distance calculated on the feature embeddings (f_{vi}, f_{vj}) defined as

$$d(f_{vi}, f_{vj}) = \|f_{vi} - f_{vj}\|_2 \tag{8}$$

Interestingly, the contrastive loss embedding l_c aims for decreasing the distance between intra class paired features and penalizes the inter class features by a margin δ as given by the contrastive cost function in Eq. (5). Specifically, a more effective loss function is shown to leverage its properties in the form of triplet loss when compared to contrastive loss.

In this manuscript, we propose to apply the triplet loss embeddings to the ranked multi-view paired features. Specified a set of multi-view training data $S = \{X_{v(i)}, y_i\} \forall i = 1\ to\ C, v = 1\ to\ N$with N views and C classes, the multi-view DML classifier focuses on learning a mapping function relating the video views $X_{v(i)}$ to $y_i$such that the predicted label $\hat{y}_i \rightarrow y_i$. In this work, we try to learn this mapping by reducing the view-specific triplet loss and the global cross-entropy loss functional. As proposed earlier, we trained a deep model D_{pp} for extracting within class maximally distant positive feature pairs $f_{v(i)} \in R^d$in d dimensions being represented as

$$f_{v(i)} = D_{pp}(X_{v(i)}, \theta_{pp}) \forall v = 1\ to\ N \tag{9}$$

Here θ_{pp} consists of trained parameters of the model D_{pp} that extracts the positive pairs within classes. This pair of features that are maximally distant within a class of views are considered as anchor positive $\{f_{a(i)}, f_{p(i)}\}$ pairs. The remaining samples of the D_{pp} were grouped into support negatives $\{f_{n(i)}\}$. This process is computed on the entire dataset. Finally, during each iteration a single triplet pair $t_z = (f_{a(i)}, f_{p(i)}, f_{n(i)})$ is constructed by following the condition $y_a = y_p \neq y_n$. Figure 4 shows the deep network used for learning from t_z. The RMVDML network learns the view mapping function through view-specific loss computed on the feature embedding space t_z. The triplet loss functional $l_{triplet}$ is

$$l_{triplet}(t_z) = \sum_{\forall N} h\left(\delta - \left\|f_{a(i)}^z - f_{n(i)}^z\right\|^2 + \left\|f_{a(i)}^z - f_{p(i)}^z\right\|\right) \tag{10}$$

where δ is the allowable margin that marks the boundary to discriminate between positive and negative pairs. Here $h(\quad) = max(, 0)$ is the hinge loss. The triplet loss aims to rationalize the weight vectors in the direction dictated by maximizing the metrics between negative pairs and minimizing metrics between positive pairs, respectively. The negative pairs are interring class features, and the positive pairs are class multi view features.

As it can observed the proposed RMVDML accessed the deep network only three times during training as compared to other multi-view multi stream networks. This has greatly reduced the load on the network trainable parameters and increased the throughput. Consequently, the triplet loss $l_{triplet}$ is used to update the weights during each iteration to penalize the negative set. For classification tasks, we need a global loss function to discriminate the classes with the help of SoftMax layers. The class label prediction is computed on the embedding space using the cross-entropy loss functional as

$$l_{Cro-Ent} = -\sum_{i=1}^{C}(y_i\, log(\hat{y}_i) + (1 - y_i)\, log(1 - \hat{y}_i)) \tag{11}$$

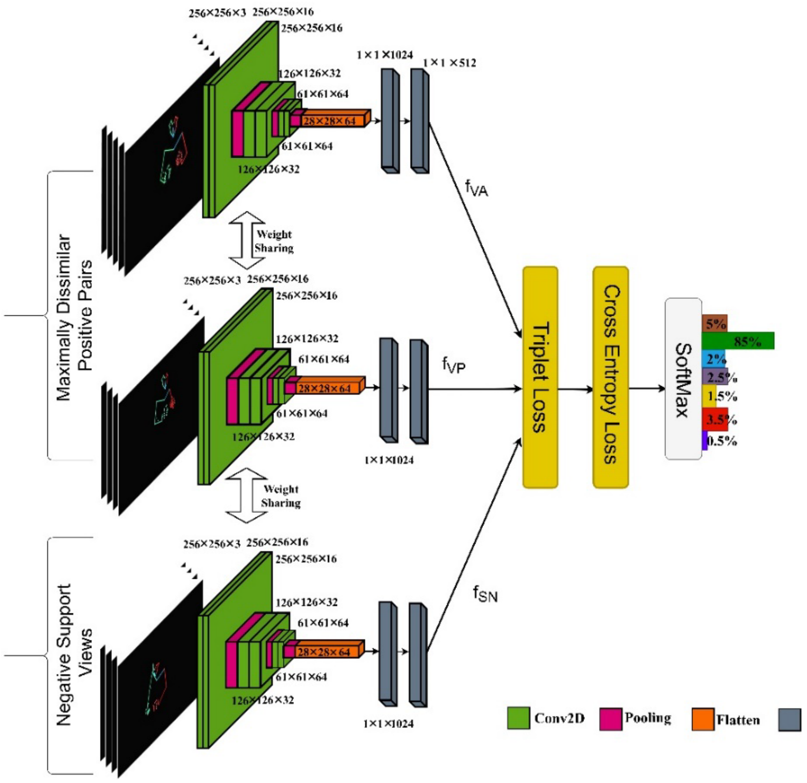

Fig. 4. Ranked multi-view deep metric leaning (RMVDML) architecture.

The RMVDML model in Fig. 4 is jointly minimized by applying the following loss function

$$l_{rmvdml} = l_{Cro-Ent} + \lambda l_{triplet} \tag{12}$$

where, λ is the loss balancing parameter between the classification loss and the view specific triplet loss embeddings.

The following procedure is instigated to train the RMVDML. First, the positive view pairing network in Fig. 3 is trained on the entire dataset to extract positive and negative pairing samples in each class. Secondly, the network in Fig. 4 is first trained with $\lambda = 0$, using only the cross-entropy loss. Next the dense+SoftMax layers are trained with $l_{triplet}$ loss function on the feature embeddings. Finally, we fine tune the whole network by selecting a value of λ on l_{rmvdml} loss. After multiple iterations, for our multi view skeletal video sign language dataset, we fixed $\lambda = 0.35$ and $\delta = 0.28$. We trained the networks with a video frame size of $256 \times 256 \times 3$. Each view consisted of 50 frames. A learning rate of 0.0001 is selected initially, which was the progressively regularized with a decay of 0.1 whenever the error became constant. The model was trained on stochastic gradient descent optimizer on an 8GB NVDIA 1070x GPU with a memory support of 16GB. The batch size of 32 was considered for training the network. The entire model has been developed in Tensorflow-2.3 with keras wrapper. Subsequent section presents the validation of our proposed method on KLEF3DSL_2Dskeletal multi view skeletal video dataset and other benchmarks.

4. Results and Discussion

RMVDML model is being evaluated on multi-view skeletal video datasets from sign language and action instances. We present an extensive description of the multi-view skeletal datasets with various training and testing ratios for evaluation. Subsequently, we present the evaluation metrics for validating the model's capabilities. Next, the proposed loss embeddings against the previously proposed models. Finally, we present a comparison of various networks against the proposed network.

4.1. Skeletal video datasets and evaluation metrics

The multi-view sign language dataset KLEF3DSL_2Dskeletal with $N = 15$views, 200 classes are generated at KL Biomechanics and Vision Computing Research Centre using 3D motion capture technology [12]. Further, the proposed model is evaluated on multi-view benchmark skeletal action datasets such as NTU RGB-D [13], SBU Kinect Interaction [14], KLYoga3D [15] and KL3D_MVaction [16]. A small subset of a data sample from KLEF3DSL_2Dskeletal is presented in Fig. 5 for a sign basketball. In this work, we are limiting our views to 15 due to computational constraints. The training testing ratios are kept constant across all networks and datasets. The selected train test ratios are 4:11, 5:10, 10:5 and 12:3. The remaining views were also evaluated but are not presented here as they have not produced any noticeable performance changes when compared to the selected ones. Since there are no multi-view sign language datasets, we evaluated our model on multi-view benchmark action datasets. Since there is a competition among the models, we selected only 40 action classes for training with 15 views from each class. The unavailability of views has prompted us to generate random views by altering the viewing angles of joints. Here, the evaluation is performed independent of the type of view in which the action is recorded. Figure 6 shows samples from the NTU RGB-D dataset. Figure 7 shows samples from KL3D_MVaction and Fig. 8 shows multi-view samples from the KLYoga3D dataset. We used mean recognition accuracy (mRA) and mean f1 score (mf1) along with precision recall curves for evaluation. All the experiments were conducted on an 8GB GPU from NVIDIA RTX 1080, 16GB DDR4 RAM and 256GB SSD.

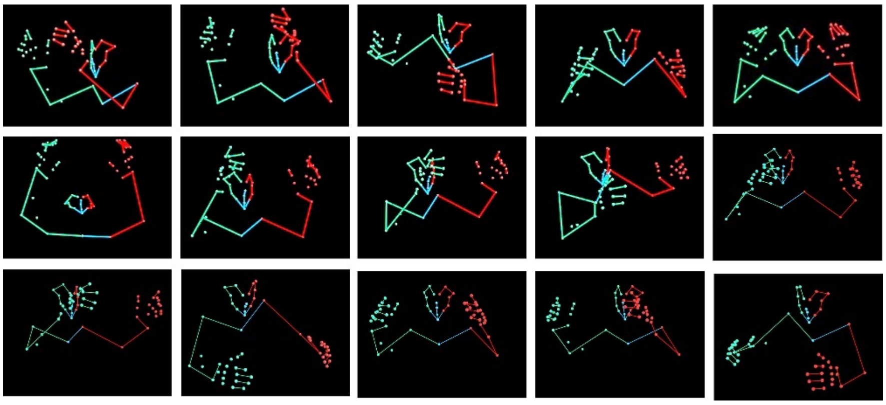

Fig. 5. KLEF3DSL_2Dskeletal sign language video dataset. A sample frame in 15 different views for the sign "Basketball".

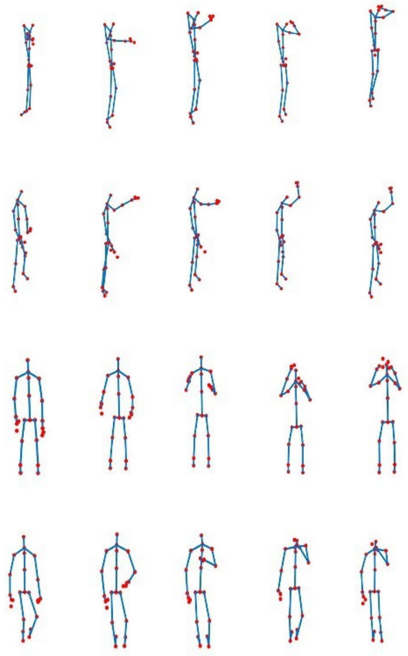

Fig. 6. NTU RGB-D in 5 views.

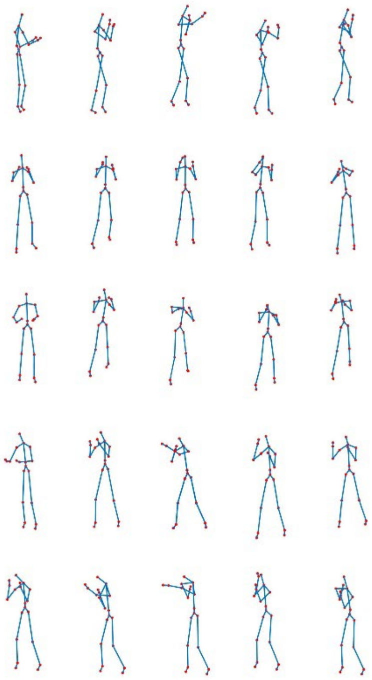

Fig. 7. KL3D_MVaction in 5 views.

4.2. Evaluating RMVDML on KLEF3DSL_2Dskeletal sign dataset

We first pass the entire training samples through the PVS network and find nonmatching pairs across each class. The view positive pairs are ranked in the order of maximum dissimilarity between views using the cosine distance function on the features from CNN in Fig. 3. The ranked positive pairs are grouped into positive pairs from each class and the remaining are grouped to form negative pairs with other class view samples. Once the grouping process is completed the positive pairs will have within class view samples and the negative set will have across the class view samples. Hence, there will be two views per positive pair and the any view other than the considered class as negative pair. Figure 9 shows the pairing process used in pre-processing stage.

The green group corresponds to the views from within class samples and the red is the negative pair formed across class view samples. During training, only one positive pair and all the other views in the negative pair are trained together. After each training episode, i.e., for each negative sample, the loss is calculated and averaged across all negative samples. Consequently, this operation is performed for the top 4 positive pairs selected by the architecture in Fig. 3 for all negative pairs in each episode. The cumulative loss on all pairs is averaged across positive pairs and then used to update the weights of the dense layers of the network in Fig. 4. This process ensures that all positive view pairs are learned against the negative view pairs. Contrasting to the traditional triplet loss models where each view has to form a pair, the P-VSN enables that only the contributing views selected will be processed which reduces the computational complexity of the network. Further, each epoch will have 4 episodes. Specifically, this process happens in the SoftMax

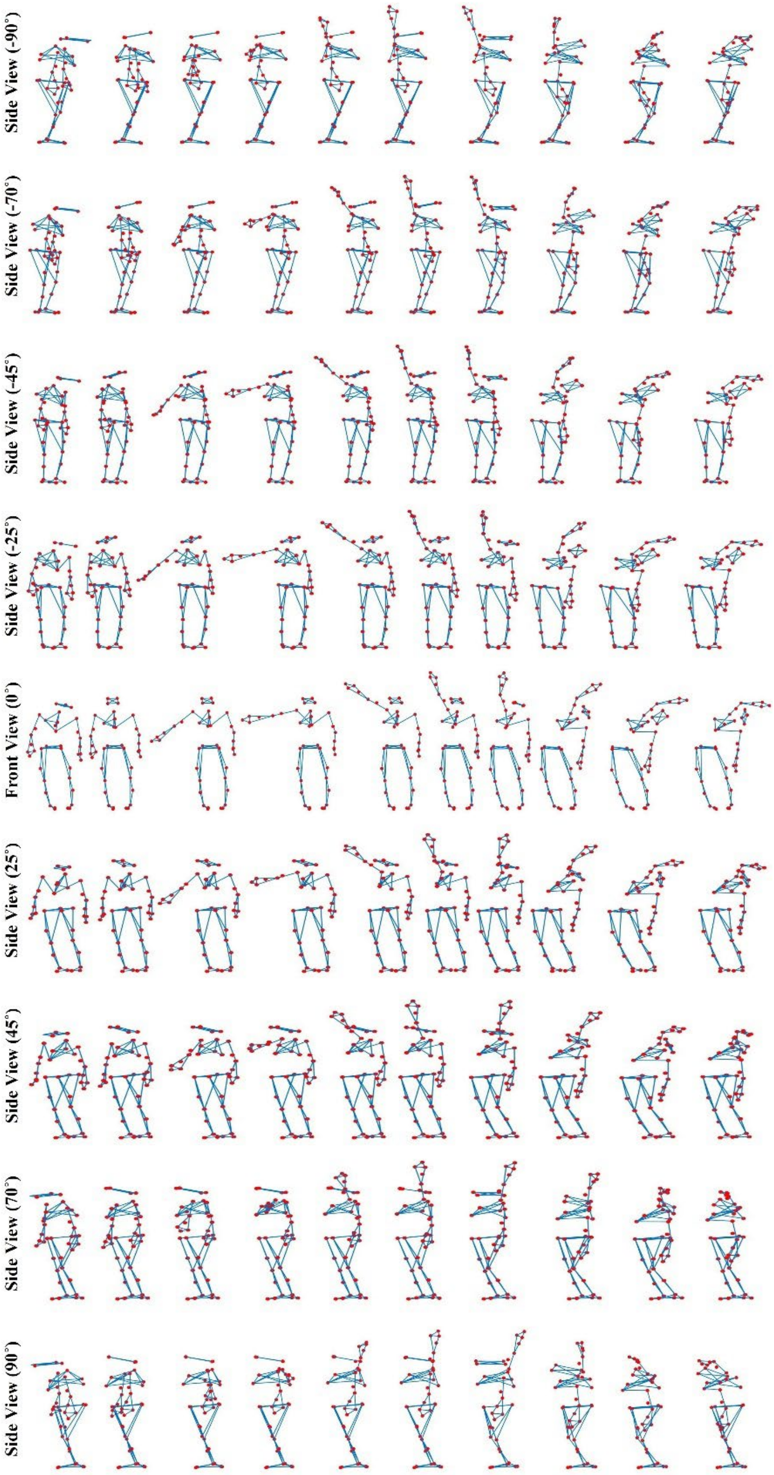

Fig. 8. KLYoga3D in 9 views.

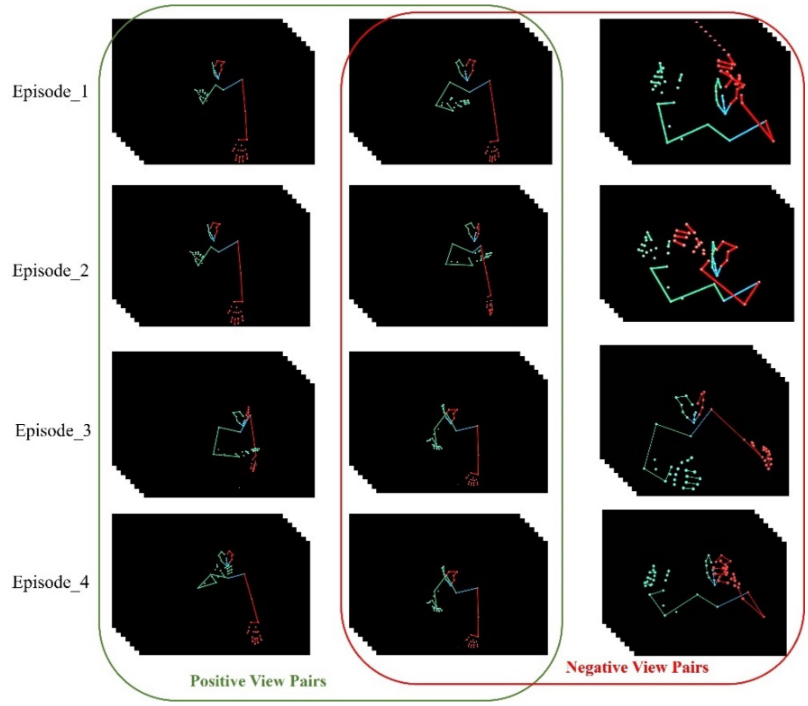

Fig. 9. The pairing structure in the proposed method using the network in Fig. 3.

and dense layers after the feature maps were built using the cross-entropy loss. Finally, the entire network is fined tuned using both the losses. We performed a 5-fold cross validation with a batch size of 16 and the accuracies were averaged over the entire dataset. For our KLEF3DSL_2Dskeletal sign language vide dataset, we present in Figs. 10 and 11, the confusion matrices for two train test ratios, 5:10 and 10:5, respectively.

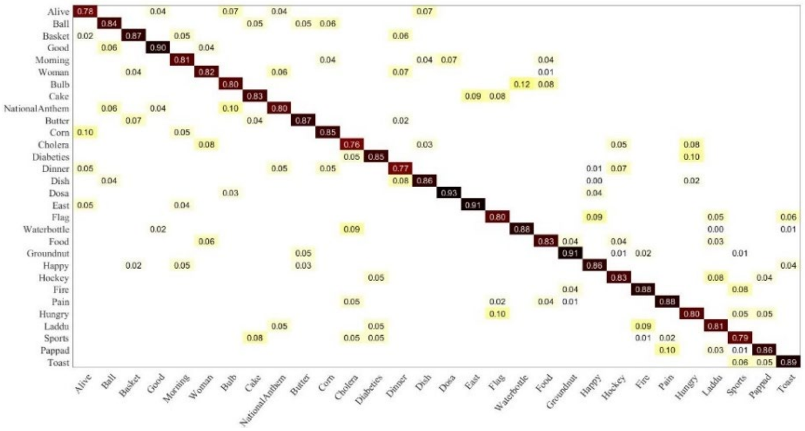

Fig. 10. Confusion matrix for 5:10 train test ratio on RMVDML model on 30 classes.

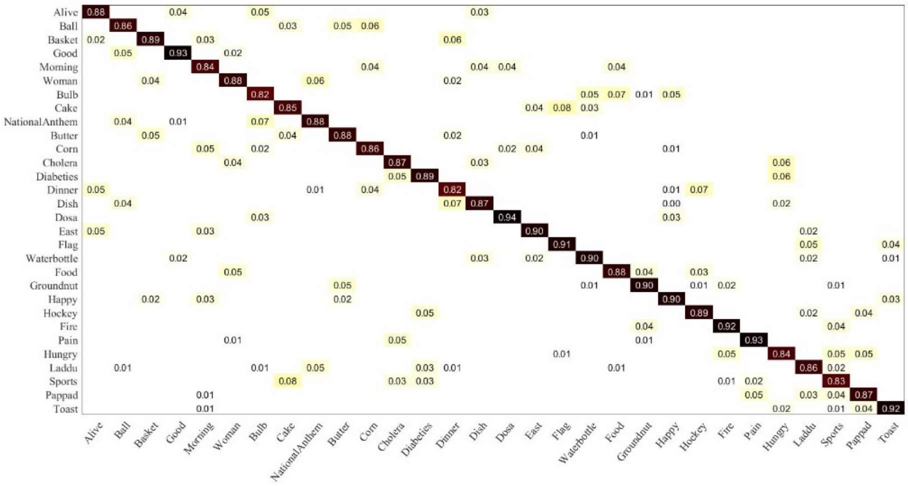

Fig. 11. Confusion matrix on train test ratio of 10:5.

The mean average recognition (mRA) obtained is around 78.34% which is far better than the traditional models such as VGG-16, ResNet-50, Google Net architectures. Similarly, mf1 is around 0.7685 for an average across a 5-fold cross validation. The λ value selected is 0.35 with a learning rate of 0.001. Successively, a comparative analysis of the proposed triplet loss embedding behavior on multiple standard networks such as VGG-16, Inception V4, GoogleNet and Resnet-50 is performed. We record the mRA and mf1 scores to indicate the performance of each of the networks against our proposed model across all datasets.

4.3. Comparison with standard architectures on proposed loss embeddings

The comparisons among state-of-the-arts against RMVDML are presented in table 1. Here it shows the effect of the loss embedding parameter λ and the effect of triplet loss on the training process. Table 1 comparison is performed on sign language skeletal video dataset KLEF3DSL_2Dskeletal. The comparative networks were trained with the proposed loss embedding function in Eq. (12). The outputs of P-VSN are used to train the standard models.

Table 1. Comparison between standard CNN models against the proposed RMVDML on KLEF3DSL_2Dskeletal sign language dataset.

Models	**mRA**	**mf1**	**mRA**	**mf1**	**mRA**	**mf1**	**mRA**	**mf1**
Loss selection parameter	$\lambda = 0.35$				$\lambda = 0$			
Train Test ratios	**5:10**		**10:5**		**5:10**		**10:5**	
VGG-16	0.709	0.707	0.728	0.727	0.610	0.591	0.636	0.602
Inception V4	0.722	0.715	0.742	0.735	0.640	0.623	0.651	0.638
GoogleNet	0.734	0.728	0.746	0.738	0.649	0.618	0.659	0.639
ResNet-50	0.683	0.693	0.702	0.692	0.581	0.526	0.618	0.602
RMVDML	0.752	0.741	0.772	0.754	0.659	0.639	0.668	0.644

From Table 1, we see that the RMVDML has been shown to outperform the standard models. This higher performance can be attributed to lesser layers which have resulted in faster training and a small number of training parameters. Our proposed network trains faster and no regularizations on weights are required as there are no vanishing gradients problems.

To validate our proposed method across cross data platforms, we apply our model to multi-view benchmark action datasets. The obtained results are compared against the standard networks. Training and testing of the networks have been uniform across all datasets. Here only 40 classes are trained and tested in 15 views. Tables 2 to 5 give the results of the experimentation.

The proposed ensemble loss gives better performance over the cross-entropy loss in all networks across all skeletal action video datasets. In the following section we evaluate the networks performance on the efficiency to retrieve class labels.

Table 2. Comparison between standard CNN models against the proposed RMVDML on NTU RGB – D action dataset.

Models	mRA	mf1	mRA	mf1	mRA	mf1	mRA	mf1
Loss selection parameter	$\lambda = 0.35$				$\lambda = 0$			
Train Test ratios	**5:10**		**10:5**		**5:10**		**10:5**	
VGG-16	0.729	0.711	0.745	0.738	0.643	0.623	0.661	0.632
Inception V4	0.742	0.724	0.787	0.767	0.676	0.664	0.698	0.672
GoogleNet	0.754	0.733	0.785	0.756	0.679	0.647	0.696	0.678
ResNet-50	0.713	0.701	0.739	0.721	0.616	0.597	0.636	0.611
RMVDML	0.762	0.752	0.793	0.768	0.683	0.651	0.697	0.679

Table 3. Comparison between standard CNN models against the proposed RMVDML on SBU Kinect Interaction dataset.

Models	mRA	mf1	mRA	mf1	mRA	mf1	mRA	mf1
Loss selection parameter	$\lambda = 0.35$				$\lambda = 0$			
Train Test ratios	**5:10**		**10:5**		**5:10**		**10:5**	
VGG-16	0.599	0.544	0.639	0.609	0.583	0.563	0.620	0.565
Inception V4	0.574	0.537	0.625	0.583	0.564	0.573	0.595	0.558
GoogleNet	0.613	0.572	0.667	0.614	0.633	0.623	0.633	0.592
ResNet-50	0.615	0.580	0.652	0.625	0.638	0.625	0.635	0.600
RMVDML	0.625	0.593	0.655	0.625	0.672	0.642	0.646	0.613

Table 4. Comparison between standard CNN models against the proposed RMVDML on KLYoga3D yoga dataset.

Models	mRA	mf1	mRA	mf1	mRA	mf1	mRA	mf1
Loss selection parameter	$\lambda = 0.35$				$\lambda = 0$			
Train Test ratios	**5:10**		**10:5**		**5:10**		**10:5**	
VGG-16	0.678	0.648	0.795	0.693	0.622	0.602	0.660	0.630
Inception V4	0.664	0.622	0.769	0.660	0.603	0.612	0.646	0.604
GoogleNet	0.706	0.653	0.790	0.695	0.672	0.662	0.688	0.635
ResNet-50	0.691	0.664	0.799	0.702	0.677	0.664	0.673	0.646
RMVDML	0.694	0.664	0.818	0.726	0.711	0.681	0.676	0.646

Table 5. Comparison between standard CNN models against the proposed RMVDML on KL3D_MVaction action dataset.

Models	mRA	mf1	mRA	mf1	mRA	mf1	mRA	mf1
Loss selection parameter	$\lambda = 0.35$				$\lambda = 0$			
Train Test ratios	**5:10**		**10:5**		**5:10**		**10:5**	
VGG-16	0.655	0.635	0.727	0.712	0.629	0.610	0.716	0.697
Inception V4	0.636	0.646	0.693	0.697	0.610	0.620	0.717	0.687
GoogleNet	0.705	0.695	0.728	0.740	0.679	0.669	0.720	0.698
ResNet-50	0.710	0.697	0.735	0.724	0.684	0.671	0.735	0.709
RMVDML	0.744	0.715	0.760	0.728	0.718	0.689	0.778	0.728

4.4. Efficiency of the proposed ensemble loss embeddings.

This section evaluates the capabilities of deep networks to retrieve information on skeletal video datasets. The evaluation is conducted by plotting precision recall curves on various datasets. The curves are obtained during testing of the networks on the proposed ensemble loss embeddings with $\lambda = 0.35$. The curves show the capabilities of the networks to retrieve different kinds of multi - view skeletal data. Figure 12 shows the plots of precision recall on the five multi - view skeletal datasets.

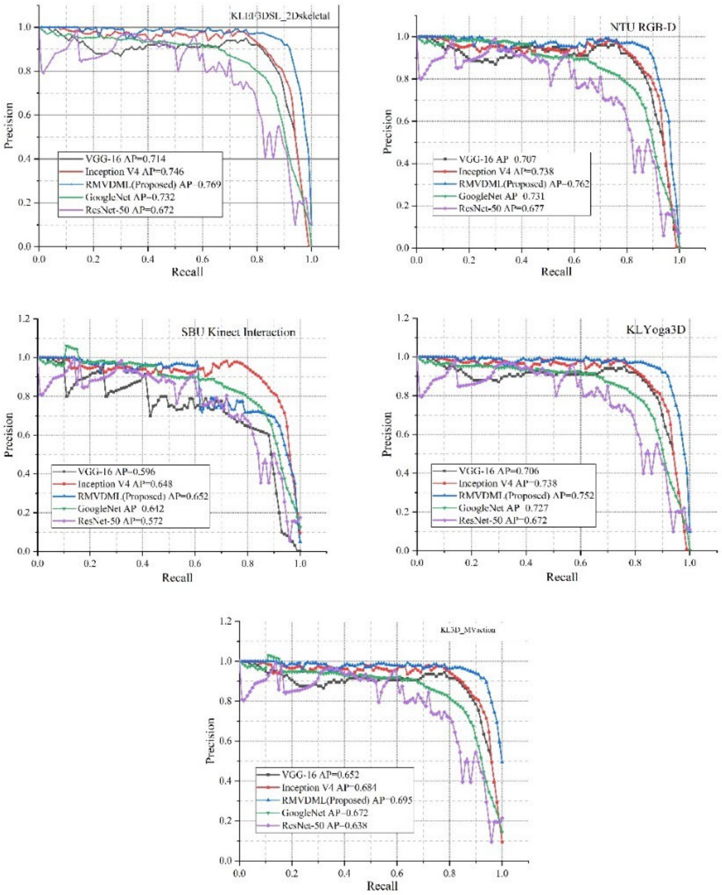

Fig. 12. Precision recall curves on various benchmark datasets.

Figure 12 shows that the proposed RMVDML has exceptional good confidence in retrieving skeletal video data when compared to other standard networks. In the next phase, we check the ensemble loss function proposed in this work against the existing losses on different networks.

4.5. Evaluating the Proposed Ensemble loss Functional.

This section presents the experiments for testing different types of losses for weight training across the baseline networks. Here, we test the sensitivity of the networks to learn across multiple loss embeddings in metric learning. Figure 13 shows the plots of the percentage of test views applied to the percentage of error generated. The plots point to the fact that the proposed ensemble loss embedding function with triplet and cross-entropy losses has the highest learnable capacity on all networks across all datasets. The cross-entropy loss has shown to have the least learning capacity. The triplet loss is the next best embeddings on the feature space.

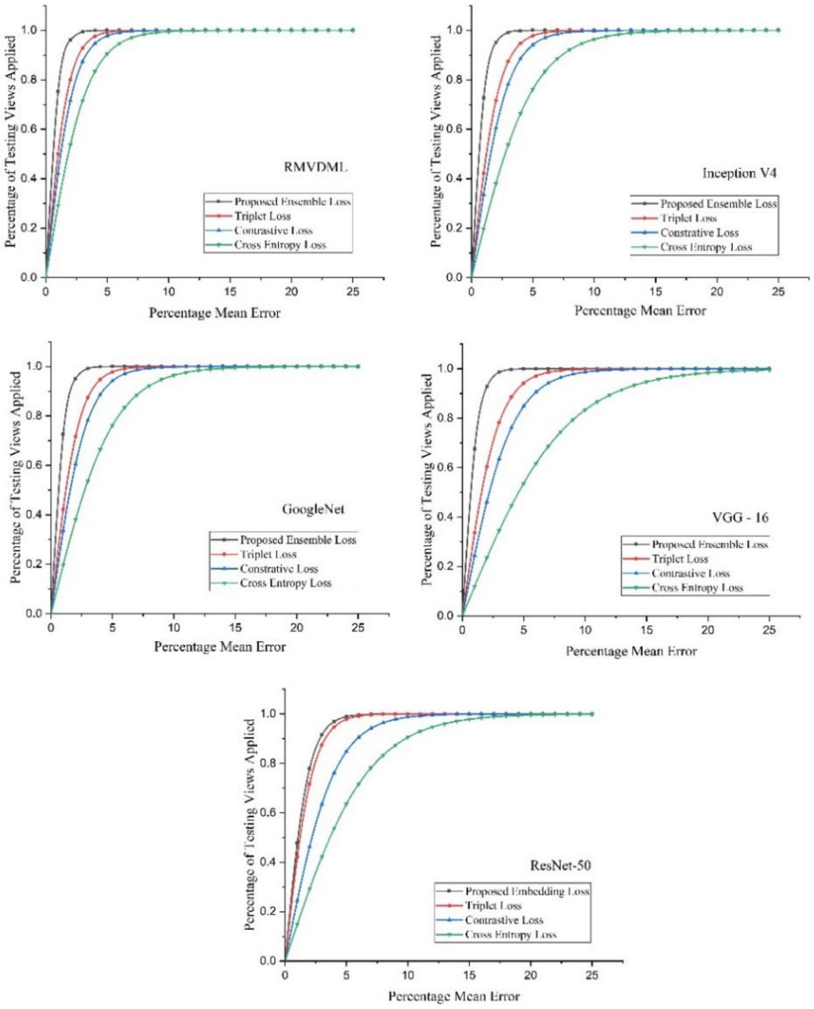

Fig. 13. Testing views fraction vs. mean error across different loss embeddings.

5. Conclusions

The work presents an ensemble loss embedding on feature space in deep metric learning model. The proposed loss is a mixture of triplet and cross – entropy loss functions. The proposed RMVDML method is a multi-view metric learning architecture for skeletal sign language recognition problems. The RMVDML is a combination of two networks, where the primary network P-VSN is a view select network and the secondary is a deep network that learns multiple views using the proposed loss embeddings. The results showed that the proposed model offers higher recognition accuracies over the standard networks across skeletal video datasets.

Nomenclatures

C	Class Number
D_{pp}	Deep metric Leaning model
F	Feature Embedding space
F_a	Anchor Features
F_p	Positive Features
F_n	Negative Features
f	Feature vectors
h	Hinge loss
l_c	Constative Loss
$l_{Cor-Ent}$	Cross Entropy Loss
l_{rmvdml}	Ensemble loss
$l_{triplet}$	Triplet loss
S	Training Samples
y	Output Labels

Greek Symbols

λ	loss balancing parameter**.**
δ	Metric Learning Hyperparameter
θ_{pp}	Trainable Parameters of DML
θ_{p-vsn}	Trainable Parameters of view select network.
X	Input vectors

Abbreviations

CNN	Convolutional Neural Networks
DML	Deep Metric Learning
LSTM	Long Short-Term Memory
mf1	mean F1 Score
mRA	mean Recognition Accuracy
P-VSN	Positive View Select Network
RMVDML	Ranked Multi View Deep Metric Learning

References

1. Kumar, E.K.; Kishore, P.V.V.; Sastry, A.S.C.S.; Kumar, M.T.K.; and Kumar, D.A. (2018). Training CNNs for 3-D sign language recognition with color

texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5), 645-649.

2. Ravi, S.; Suman, M.; Kishore, P.V.V.; Kumar, M.T.K.; and Kumar, D.A. (2019). Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition. *Journal of Computer Languages*, 52, 88-102.
3. Cui, R.; Liu, H.; and Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7), 1880-1891.
4. Guo, D.; Zhou, W.; Li, H.; and Wang, M. (2018). Hierarchical LSTM for sign language translation. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 32(1), 6845-6852.
5. Mittal, A.; Kumar, P.; Roy, P.P.; Balasubramanian, R.; and Chaudhuri, B.B. (2019). A modified LSTM model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, 19(16), 7056-7063.
6. Ge, W.; Huang, W.; Dong, D.; and Scott, M.R. (2018). Deep metric learning with hierarchical triplet loss. *Proceedings of the European Conference on Computer Vision* (*ECCV* 2018), Munich, Germany, 272-288.
7. Wang; J.; Wang, K.-C.; Law, M.T.; Rudzicz, F.; and Brudno, M. (2019). Centroid-based deep metric learning for speaker recognition. *Proceedings of the ICASSP* 2019 - 2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, United Kingdom, 3652-3656.
8. Hu, J.; Lu, J.; and Tan, Y.-P. (2014). Discriminative deep metric learning for face verification in the wild. *Proceedings of the* 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 1875-1882.
9. Gutoski, M.; Lazzaretti, A.E.; and Lopes, H.S. (2020). Deep metric learning for open-set human action recognition in videos. *Neural Computing and Applications*, 33, 1207-1220.
10. Kang, J.; Fernandez-Beltran, R.; Ye, Z.; Tong, X.; Ghamisi, P.; and Plaza, A. (2020). Deep metric learning based on scalable neighborhood components for remote sensing scene characterization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12), 8905-8918.
11. Yi, D.; Lei, Z.; Liao, S.; and Li, S.Z. (2014). Deep metric learning for person re-identification. *Proceedings of the* 2014 22*nd International Conference on Pattern Recognition*, Stockholm, Sweden, 34-39.
12. Kishore, P.V.V.; D. Kumar, D.A.; Sastry, A.S.C.S.; and Kumar, E.K. (2018). Motionlets matching with adaptive kernels for 3-d Indian sign language recognition. *IEEE Sensors Journal*, 18(8), 3327-3337.
13. Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. (2016). NTU RGB+D: A large scale dataset for 3d human activity analysis. *Proceedings of the* 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 1010-1019.
14. Li, M.; and Leung, H. (2016). Multiview skeletal interaction recognition using active joint interaction graph. *IEEE Transactions on Multimedia*, 18(11), 2293-2302.

15. Maddala, T.K.K.; Kishore, P.V.V.; Eepuri, K.K.; and Dande, A.K. (2019). YogaNet: 3-D yoga asana recognition using joint angular displacement maps with ConvNets. *IEEE Transactions on Multimedia*, 21(10), 2492-2503.
16. Srihari, D.; Kishore, P.V.V.; Kumar, E.K.; Kumar, D.A.; Kumar, M.T.K.; Prasad, M.V.D.; and Prasad, C.R. (2020). A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-D data. *Multimedia Tools and Applications*,79(17-18), 11723-11746.
17. Jiang, X.; Satapathy, S.C.; Yang, L.; Wang, S.-H.; and Zhang, Y.-D. (2020). A Survey on artificial intelligence in Chinese sign language recognition. *Arabian Journal for Science and Engineering*, 45(12), 9859-9894.
18. Chong, T.-W.; and Kim, B.-J. (2020). American sign language recognition system using wearable sensors with deep learning approach. *The Journal of the Korea institute of Electronic Communication Sciences*, 15(2), 291-298.
19. Aly, S.; and Aly, W. (2020). DeepArSLR: A novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition. *IEEE Access*, 8, 83199-83212.
20. Xiao, Q.; Qin, M.; and Yin, Y. (2020). Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, 125, 41-55.
21. Li, D.; Opazo, C.R.; Yu, X.; and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proceedings of the* 2020 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Snowmass, CO, USA, 1448-1458.
22. Huang, J.; Zhou, W.; Li, H.; and Li, W. (2018). Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9), 2822-2832.
23. Imran, J.; and Raman, B. (2020). Deep motion templates and extreme learning machine for sign language recognition. *The Visual Computer*, 36(6), 1233-1246.
24. Kumar, P.; Gauba, H.; Roy, P.P.; and Dogra, D.P. (2017). Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86, 1-8.
25. Pigou, L.; Dieleman, S.; Kindermans, P.-J.; and Schrauwen, B. (2014). Sign language recognition using convolutional neural networks. *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, 572-578.
26. Aloysius, N.; and Geetha, M. (2020). Understanding vision-based continuous sign language recognition. *Multimedia Tools and Applications*, 79(31), 22177-22209.
27. Xing, Y.; and Zhu, J. (2021). Deep learning-based action recognition with 3D skeleton: A survey. *CAAI Transactions on Intelligence Technology*, 6(1), 80-92.
28. Hussain, T.; Muhammad, K.; Ding, W.; Lloret, J.; Baik, S.W.; and de Albuquerque, V.H.C. (2021)(in press). A comprehensive survey of multi-view video summarization. *Pattern Recognition*, 109, 107567.
29. Gao, Z.; Zhang, H.; Xu, G.P.; Xue, Y.B.; and Hauptmann, A.G. (2015). Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. *Signal Processing*, 112, 83-97.

30. Zheng, J.; Jiang, Z.; and Chellappa, R. (2016). Cross-view action recognition via transferable dictionary learning. *IEEE Transactions on Image Processing*, 25(6), 2542-2556.
31. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; and Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *Proceedings of the* 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 2136-2145.
32. Ullah, A.; Muhammad, K.; Hussain, T.; and Baik, S.W. (2021). Conflux LSTMs network: A novel approach for multi-view action recognition. *Neurocomputing*, 435, 321-329.
33. Si, C.; Chen, W.; Wang, W.; Wang, L.; and Tan, T. (2019). An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, California, 1227-1236.
34. Wang, D.; Ouyang, W.; Li, W.; and Xu, D. (2018). Dividing and aggregating network for multi-view action recognition. *Proceedings of the European Conference on Computer Vision* (*ECCV* 2018), Munich, Germany, 457-473.
35. Kishore, P.V.V.; Prasad, M.V.D.; Raghava Prasad, Ch.; and Rahul, R. (2015). 4-Camera model for sign language recognition using elliptical Fourier descriptors and ANN. *Proceedings of the* 2015 *International Conference on Signal Processing and Communication Engineering Systems*, Guntur, India, 34-38.
36. Kaya, M.; and Bilge, H.Ş. (2019). Deep metric learning: A survey. *Symmetry*, 11(9), 1066, 1-26.
37. Shorfuzzaman, M.; and Hossain, M.S. (2021). MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. *Pattern Recognition*, 113, 107700.
38. Hoffer, E.; and Ailon, N. (2015). Deep metric learning using triplet network. *Proceedings of the International Workshop on Similarity-Based Pattern Recognition*, Copenhagen, Denmark, 84-92.
39. Zheng, W.; Lu, J.; and Zhou, J. (2021). Hardness-aware deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9), 3214-3228.
40. Wang, J.; Zhou, F.; Wen, S.; Liu, X.; and Lin, Y. (2017). Deep metric learning with angular loss. *Proceedings of the* 2017 *IEEE International Conference on Computer Vision ICCV* 2017, Venice, Italy, 2593-2601.
41. Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Proceedings of the* 30*th International Conference on Advances in Neural Information Processing Systems* 29 (*NIPS* 2016), Barcelona, Spain, 1-9.