Received: December 16, 2022. Revised: February 1, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Meta Triplet Learning for Multiview Sign Language Recognition

Suneetha Mopidevi[1,2] **Prasad M.V.D**[1] **Venkata Vijay Kishore Polurie**[1]
Anil Kumar Dande[3]*

[1]*Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522302, India*
[2]*Department of Electronics and Communication Engineering,*
Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, India
[3]*Department of Electronics and Communication Engineering, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, 523272, India*
*Corresponding author's Email: danilmurali@gmail.com

Abstract: Multiview video processing for recognition is a hard problem if the subject is in continuous motion. Especially the problem becomes even tougher when the subject in question is a human being and the actions to be recognized from the video data are sign language. Although many deep learning models have been successfully applied for sign language recognition (SLR), very few have considered multiple views in their training set. In this work, we propose to apply meta metric learning for video-based sign language recognition. Contrasting to traditional metric learning where the triplet loss is constructed on the sample-based distances, the meta metric learns on the set-based distances. Consequently, we construct meta cells on the entire multiview dataset and perform a task-based learning approach with respect to support cells and query sets. Additionally, we propose a maximum view pooled distance on sub-tasks for binding intraclass views. The results of experiments conducted on the multiview sign language dataset and four action datasets show that the proposed multiview meta metric learning model (MVDMML) achieves 11% higher performance than the baselines.

Keywords: Deep meta metric learning, Multiview sign language and action recognition, Triplet loss.

1. Introduction

Sign language recognition (SLR) has been explored in multiple signalling environments such as 1/2/3D over the past three decades. Despite its recent successes with deep neural networks (DNNs) such as convolutional neural networks (CNNs) [1-3] and long shot term memory networks (LSTMs) [4, 5], the development of a deployable SLR is far from reality. This is due to problems relating to subjects hand and body movements that will augment the single-view data into a multi view data recovery problem. This kind of multi view problems is 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. These problems were addressed by applying global optimization models as shown in Fig. 1(a) for multi view learning. The resulting outputs from global optimization methods on entire training sets has shown to overfit on the common features between views. During testing, the within class query view samples were found to linearize towards the common features in other classes providing ambiguous outputs from the SoftMax layer. Unlike the above methods that are applied on the entire dataset, metric learning focuses on pairs of samples from the training set to maximize the distance between the samples of the training set.

The goal of this work is to develop a view sensitive sign language recognition system using

Received: December 16, 2022. Revised: February 1, 2023.

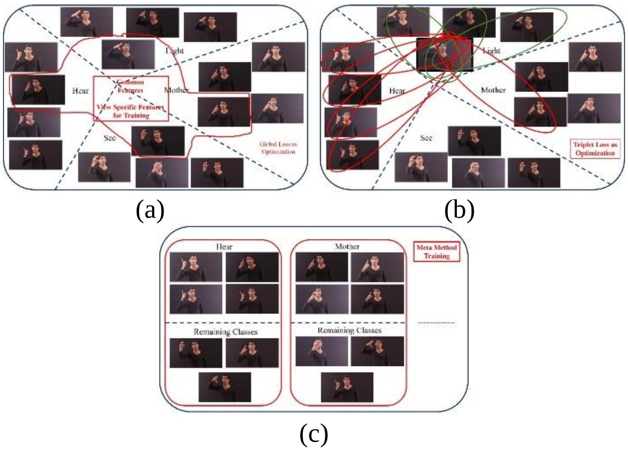


Figure. 1 A comparison between global optimization methods: (a), triplet loss metric learning model, (b) and our proposed meta metric learning, and (c) for multi view video data recognition with sign language

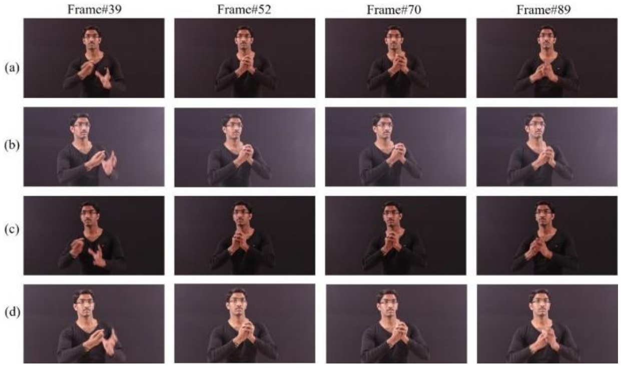


Figure. 2 The multi view sign language dataset, KL_MV2DSL. It has 4 views from different directions and one view from frontal direction with a total of 5 views per class

model of deep metric learning [6]. Generally, deep metric learning has been a successful algorithm in the fields of speaker identification [7], face recognition [8], action recognition [9], satellite image classification [10] and person re-identification [11] problems. However, this is the first instance where deep metric learning is being investigated for multi view recognition problems. In DML the model is trained to learn the similarities between the intra class variables and discriminate their inter class dissimilarities. The learning process is instigated using a loss function defined by contrastive or triplet loss. The triplet loss embedding is a distance metric that tries to maximize the gap between intra and inter class features against an anchor image, whereas the contrastive loss does the same thing without the anchor.

However, the implementation of the multi view metric learning architecture is challenging and computationally expensive due to multiple combinations of anchor positive and anchor negative pairs as shown in Fig. 1(b). For example, in a c class, m view dataset with f frames per view, we can have a total of $m \times f$ frames per class. The total number of frames across all classes in $m \times f \times c$. Let a frame in a particular view be selected as anchor, then the number of anchor positive pairs from within a class will be $(mf)C_2$. Consequently, the number of anchor negative pairs across classes is $(c-1)mfC_2$, which have thrown memory exceptions during training process. In our Multiview sign language dataset, we have 5 views by 5 different subjects in 200 classes. Each video sign has 120 frames, which makes the total dataset into a 600K frame repository. As a result, the DML algorithm with triplet loss comes under massive workloads to transform the inputs into pairs. Accordingly, the above constraint on DML on multi view data increases exponentially with the increase in number of views per label. Hence, to develop a computationally efficient environment for multi view metric learning using DML, we transform the above problem into a deep meta metric learning [12] problem as depicted in Fig. 1(c). Contrasting to the metric learning with triplet loss embeddings, the training process is sampled into sub – tasks to learn transferable information across multiple views.

This work proposed to apply deep meta metric learning [12] to multi view problem in sign language (action) recognition tasks. Experiments are conducted to justify the capabilities of the proposed model for multi view problems against different metric learning classifiers. In particular, we show the performance of the proposed meta metric model on our multi view sign language video dataset KL_MV2DSL, and other action benchmarks such as NTU RGB D[13], MuHAVi[14], WEIZMANN[15] and NUMA[16]. The fig.2 offers a glimpse of the video frames from our KL_MV2DSL, the multi view sign language dataset.

Action datasets are selected because of the unavailability of rich multi view sign language dataset. The rest of the paper is organized as follows. The section 2 describes the pros and cons of sign language recognition systems based on deep learning with an emphasis on multi view action recognition models. The proposed methods are being presented in section 3. Subsequent sub sections in 4 demonstrate the capabilities of multi view meta metric learning in recognizing multi view SL data and other human action multi view benchmark datasets. Finally, conclusions are drawn on the proposed method in section 5.

2. Background

Since the work is a mixture of many interconnected areas, we review these areas discretely instead of a single entity. We present the review of sign language, action recognition, deep metric learning and meta learning in the following subsections.

2.1 Sign language recognition

Sign language recognition has been practiced in various forms based on data, features and classification algorithms [1]. The data usually comes from 3 sources, hand gloves (1D) [17], video cameras (2D) [18] and Kinect or leap motion (3D) [19]. The 4th and unique high-priced source is motion capture technology that has produced high precision synthetic sign language skeletal data [20]. Despite costly, 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 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 actually solved some type of spatial, temporal or paired representation of video object data effectively as features. These features are further classified using all the traditional machine learning algorithms. The most popular classifiers were hidden markova 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. Multi modal video sequences that are fed into multiple streams of a CNN are predominantly researched which have shown evidence of 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 solution for real time implementation.

2.2 Multi-view action recognition

Eventually, to develop a real time SLR or HAR system, it is intuitive to initiate multi view learning. Therefore, in the last couple of years, multi view learning has taken centre stage [28]. Multi view HAR has evolved through research using dictionary learning [29], neural networks with adaptable views [30], convolutional neural networks [31] and deep attention models [32], to name a few. However, 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 [33]. 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 the variations in multi view data [25].

2.3 Deep metric learning

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 the ability to cluster highly similar within class samples by learning the loss dynamics across different classes [34]. The loss dynamics in calculated using the contrastive and triplet functions [35] which are used to train the deep networks. In the past few years, DML is applied to multiple vision-based applications such as person identification [36], face recognition in the wild [37], speaker identification [38], image classification [39] and remote sensing data [40]. Moreover, there are multiple procedures in which the loss can be included into the objective function apart from triplet and contrastive techniques. Some of them are, hierarchical triplet loss [41], hard triplet loss [42], multi similarity loss [43] and n – pair multiclass loss [44]. All the losses have distinctive advantages, especially in maximizing within class and minimizing across class similarities for maximum performance. Lastly these losses are difficult to implement due to multiple regularizations that are specific to a problem at hand.

In this work, we follow the model proposed in [12], which closely relates to methods of matching networks. The matching networks are trained to learn a set of task specific classifiers to solve the problem of few shot learning by weighing these nearest neighbour classifiers. Contrastingly, the weighing nearest neighbourhood classification in few shot learning is replaced by metric learning for visual recognition problems [12]. The work in [12] uses meta formulation of the visual classification problem with hard sample mining that has been found to be computationally efficient with good recognition accuracies. However, applying it to multiview video data recognition problems has increased the complexity of the hard sample mining, where the algorithm found it difficult to distinguish between the query and meta samples. Hence, we propose a more simplified maximum pooling distance metric for the multiview meta metric learning model. The proposed work is simpler than other multiview models in three aspects: 1. It uses a meta metric learning model for solving multiview sign language recognition. 2. Our meta metric learning model uses simplified task specific convolutional neural networks for feature extraction with 6 layers. 3. The metric used is computationally efficient and highly discriminative across datasets.

3. Methodology

This section details the process of deep meta metric learning for multi view video-based sign language (action) recognition. We develop the theory and implementation procedures for an end – to – end trainable system.

3.1 Deep meta metric learning

Here, we developed the theoretical background of deep meta metric learning from the works in [12]. Meta learning is to reuse the learned experience from across tasks in a systematic approach. In other words, it trains a deep learning model with the experience gained in other trained models across similar tasks. The entire meta learning process constitutes of two parts. In the first part, we form meta-data from previously learned tasks and models. These include, data configurations, hyperparameter settings, dataflow pipelines and model evaluations such as trained weights, biases and accuracies. In the second part, the prior meta data guides the formulated models to learn new tasks by extracting features and transferring knowledge. Contrastingly, traditional deep learning happens to miss on the prior knowledge form other tasks. However, in metric learning, the learning is approached by calculating distance metrics across similar and dissimilar pairs of data. Combining the metric with meta learning has in the past improved the performance of person re-identification problems [12]. Therefore, this work formulates a multi view sign language (action) recognition solution through deep meta metric problem.

Traditional deep learning algorithms learn the trainable parameters θ by optimizing a cost function$L$on the overall training sample observations as

$$\theta = arg\min_{\theta} L(\theta; x, y) \tag{1}$$

Where, L defines the global loss across all the training samples and (x, y) are the training pairs, with x as training data and $y$as the labels. In metric learning, triplet loss has shown to improve the deviation between the inter and intra class embeddings over the contrastive loss, which is computed on triplets (x_a, x_i, x_j) as against doublets in the later [6]. The x_a is the anchor of the triplet that has the same label as the x_i, called the positive sample and a different label with x_j, the negative sample. This will stimulate the CNNs to construct an embedding space in which the positive sample x_i is pushed close to anchor x_a and simultaneously push the negative sample x_j away from x_a. The cost function L^{trip} is defined as

$$L^{trip}(x_a, x_i, x_j, y_{ij}) = \sum_{a,i,j} max\left(\left(d(x_a, x_i) - d(x_a, x_j) + \delta\right)^2, 0\right) \tag{2}$$

Where δ is the minimum margin that separates the positive and negative samples. Triplet embedding uses one pair of positive and negative sample per iteration. Besides all the pairs are treated equally and loss is calculated on all possible pairs, which make it sometimes trivial to find hardest pairs. Also, if we have to apply this for a multi view problem with greater than two views, the pair formation becomes complex and computationally expensive.

In the proposed deep meta metric learning, we propose to apply metric learning in a meta way for multi view recognition problems. First, we divide the single cost function on the entire training observations into sub tasks. Secondly, we train a deep model to learn the meta metric on all these divided sub tasks. Accordingly, task distribution $\Delta(T)$ divides the training samples into test and all other sub tasks. Incidentally, the objective function of deep meta metric learning is formulated as

$$\theta = arg\min_{\theta} P_{T_s \sim \Delta(T)}[L_s(\theta; x_s, y_s)] \tag{3}$$

Where $L_s(\theta; x_s, y_s)$is the objective function on the sampled subtasks $T_s$and $P$is expectation on all evaluations on distributed tasks across the training samples. The evaluations used is accuracy. In particular, for a C class training set, we randomly sample $C_{New}(C_{New} < C)$classes from the Cas a new task. Subsequently, we sample randomly select support set $S = \left\{s_i^{c_{new}} | i = 1, \ldots, c_k^{c_{new}}\right\}$ and query set $Q = \left\{q_i^{c_{new}} | i = 1, \ldots, c_q^{c_{new}}\right\}$ for each of the sub task T_s. Here, $c_{new} = 1, \ldots, C_{New}$ forms a set of new classes. The number of support and query samples in each class are considered as equal $c_k^{New} = c_k$ and $c_q^{New} = c_q$. In each iteration, the metric learning is applied by computing the distance metric between the query and support samples. The overall meta metric learning cost function can be formulated as

$$\theta = arg \min_{\theta} P_{T_s \sim \Delta(T)} \left[P_{S,Q \sim T_s}[L_s(\theta; S, Q)]\right] \tag{4}$$

3.2 Multi view meta metric learning per episode

Considering all views in a class as support data points that lie on a view manifold, we define a multi view meta cell which learns through meta metric in each episode as,

$$M_V^{c_{new}} = \left\{\sum_{i=1}^{c_q^{new}} \alpha_i^{c_{new}} f\left(s_i^{c_{new}}\right) | \sum_{i=1}^{c_q^{new}} \alpha_i^{c_{new}} = 1, \alpha_i^{c_{new}} = [0,1]\right\} \tag{5}$$

Where $f(.)$ is the embedding function on the support data set that is implemented using the deep neural network with parameters θ. The coefficient $\alpha^{c_{new}} \in [0,1]$ maintains the convexity of the multi view meta cell. In conventional multi view learning, the model is trained to identify a particular query class over the entire dataset. Interestingly, multi view metric learning uses distances between view sample pairs to determine an output class label. Contrastingly, multi view meta metric learning optimizes the metrics between sets of multi view support meta cells and the query video frames. This model optimizes the with in class metrics from multiple views to learn a discriminative distance metric for classification. The set distance metric between the multi view meta cells and the query sample is computed as

$$D\left(q_j^{c_{new}'}, M_V^{c_{new}}\right) = d_j^{c_{new}} = \sum_{i=1}^{c_q^{new}} \alpha_i^{c_{new}} \left(f\left(q_j^{c_{new}'}\right), f\left(s_i^{c_{new}}\right)\right) \tag{6}$$

Where, $q_j^{c_{new}'}$ is the query sign video in the j^{th} sample view in the c_{new}' class, $M_V^{c_{new}}$ is the multi view meta cell in the class c_{new}. The $d^{c_{new}}$ denotes the distance metric between the query video and multi view meta cell or the support sample data points.

Given a query view of a certain class from the formulated query view set Q_V, the model learns to minimize the distance metric between Q_V and meta view cell of the same class and maximize the distance with other meta cell classes. Subsequently, the triplet metric learning loss function is initiated on the sampled meta triplets with $M_V\{M_V^1, \ldots., M_V^M\}$ and one Q_V sample on single positive meta cell per class $M_V^{c_{new}}\{c_{new} = c_{new}'\}$and all other class views as negative meta cells as

$$L_{Meta_{tri}}(q_j) = \sum_{c_{new} \neq c_{new}'} max\left(0, d_j^{c_{new}'} - d_j^{c_{new}} + \tau\right) \tag{7}$$

Where τ is the hyperparameter allowable gap between the positive and negative meta cell pairs. The $c_{new} \neq c_{new}'$ forms the negative meta cell pair and vice versa for positive meta cell pair. In practice, the final classification layer the predictions are computed using logistic loss function and we apply this in place of max(0, d) to limit the range of loss function in Eq. (7). The logarithmic loss embedding on the meta metric learning is formulated as

$$L_{MV\_trip}^{log(q_j)\, log\left(1+\sum_{c_{new} \neq c_{new}'} exp\left(d_j^{c_{new}'} - d_j^{c_{new}} + \tau\right)\right)} \tag{8}$$

The value of τ is selected in such way that $d_j^{c_{new}} \ll d_j^{c_{new}'}$. The proposed multi view meta learning model is visualized in Fig. 3, where the distance between the query and positive meta cell is considerably smaller than that of query and the negative meta cell pairs. The query sample is metrically close in distance with the samples in positive meta cell with a set margin when compared to all other negative meta cell views in different classes.

Eq. (8) has similarities with the regular SoftMax loss that outputs a probability estimate of closeness between the query sample q_j and meta cell M_V. Incepted from [12], the SoftMax loss is computed within the meta space by adding margins on the negative loss to the negative meta views to increase the separation between the positive and negative meta views.

The negative SoftMax loss for multi view meta learning model is

Received: December 16, 2022. Revised: February 1, 2023.

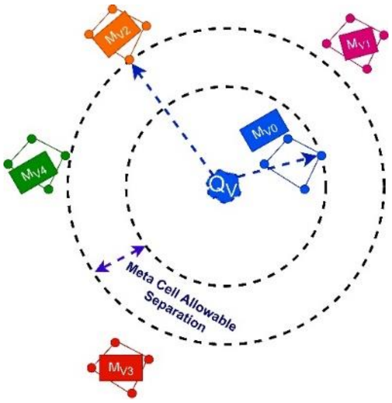


Figure. 3 Visual illustration of the proposed meta cell allowable separation used in the proposed method

$$L_{MV\_trip}\quad log(q_j)\, log\frac{exp\left(-d_j^{c'_{new}}\right)}{exp\left(-d_j^{c'_{new}}\right)+\sum_{c_{new}\neq c'_{new}} exp\left(-d_j^{c_{new}}+\tau\right)} \qquad (9)$$

Finally, the above classification loss in each episode is optimized under the influence of task distribution and data splitting into random query and support views. The formulated model under multi view meta learning model is

$$\theta = \arg\min_{\theta} P_{T_s\sim\Delta(T)}\left[P_{S,Q\sim T_s}\sum^{\Sigma}_{q_j^{c'_{new}}\in Q} L_{MV\_trip}^{log\left(q_j^{c'_{new}}\right)}\,[]\right] \qquad (10)$$

3.3 Maximum pooled meta view mining

In the proposed multi view deep meta metric learning (MVDMML) model, we replaced view-based distances with set-based distances as shown in Eq. (6). However, as pointed in [12], it was difficult to realize the generalized definition by optimizing the distance formulation. Alternatively, two solutions were proposed in [12]: the centre support distance and the other hard mining distance.

The centre support distance is computed between the averaged multi view meta cell features and the query sample which computes a point – to – point variations. The averaged multi view meta cell features represent the entire class of a meta cell. The multi view centre support distance

$$D\left(q_j^{c'_{new}}, M_V^{c_{new}}\right) = d\left(f\left(q_j^{c'_{new}}\right), \frac{1}{c_q^{new}}\sum_{i=1}^{c_q^{new}} f\left(s_i^{c_{new}}\right)\right) \qquad (11)$$

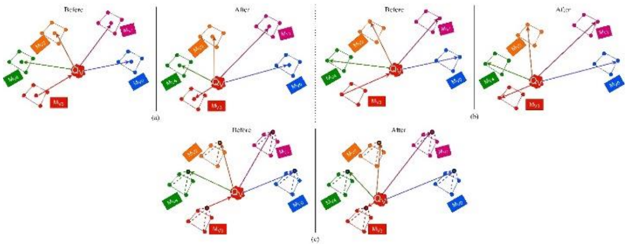


Figure. 4 Centre support distance and hard mining distance as used in [12]. (a) Centre support distance before the after execution, (b) Hard mining distance, and (c) Our proposed maximum pooled distance across multiple views, where the distance is computed between the query sample and maximum pooled features in a meta cell across multiple views

The term $\frac{1}{c_q^{new}}\sum_{i=1}^{c_q^{new}} f\left(s_i^{c_{new}}\right)$ gives the average pooling on multiple views with in the meta cell. This mode of distance computation is challenging as the query sample finds it difficult to identify hard and easy samples within the meta cells. Hence, hard sample mining is proposed in [12] which calculates the distance between hard samples and easy samples in the meta cells. The objective of hard sample mining is the maximize the distance between the query and the hard samples in inter class meta cells by simultaneously minimizing inter class distances for selection of easy samples. The hard sample mining is formulated as

$$D_{\hbar}\left(q_j^{c'_{new}}, M_V^{c_{new}}\right) = \begin{cases} \arg\max_i\left(d\left(f\left(q_j^{c'_{new}}\right), f\left(s_i^{c_{new}}\right)\right)\right)\forall c'_{new} = c_{new} \\ \arg\min_i\left(d\left(f\left(q_j^{c'_{new}}\right), f\left(s_i^{c_{new}}\right)\right)\right)\forall c'_{new} = c_{new} \end{cases} \qquad (12)$$

Instead of finding negative hard samples gradually by calculating true negatives, the solution in [12] proposes to take the hard mining process into the set-based distance to reduce intra class variations. The hard mining process has indeed increased the intra class variances by highlighting the outliers and penalizing them for being a part of the meta cells. The above two mining process are visualized in Figs. 4 (a) and (b). Figs. 4 (a) and (b) shows that the centre support distance pushes all the samples in the negative meta cells, and the hard mining process does this in a selective manner by maximizing distances with only hard samples. Specifically, the hard mining process on meta view learning has resulted in computational complexity during implementation phase. For instance, the query view needs to check against all views in the meta cell to identify the negative samples and all the samples within a meta cell has a set of unique variant features. To take advantage of these view variant features that are spatially distributed, we propose an upgraded model with view variant feature pooling mechanism. Instead of average pooling, we propose maximum pooling on meta cell views to extract the view specific features. Fig. 4 (c) shows the model for maximizing the distance between query and maximum pooled view variant features. The maximum pooled meta distance is formulated as

3.4 Training and testing

We used TensorFlow wrapped keras for implementation of the proposed meta metric multi view model. Euclidean distance metric was applied to Eq. (6) for separating the query features from the meta features. To generalize the training and testing pipelines across datasets and models, we set the standard train, validate and test ratios at 70:10:20 percent of the entire dataset. Subsequently, all the hyper parameters of the model were fixed across all experiments to attain uniformity during comparison. The number of sub classes per episode was set to $M = 16$ with $c_{new} = 4$ support samples per meta cell. This number of support samples may change based on the availability of views in the dataset. The meta cell allowable separation is selected to be $\tau = 0.22$ in our objective function (8). We used Adam optimizer with a learning rate of 0.0001 with a decay of 10% whenever it became constant for more than 10 epochs. The following section gives details of the experiments and provide an insight into the capabilities of the proposed approach.

4. Experiments, results and insights

The proposed MVDMML is trained and tested for recognizing signs from multi view video data. This section starts by describing our multi view sign language dataset, KL_MV2DSL and four other benchmark multi view action datasets NTU RGB D, MuHAVi, WEIZMANN and NUMA. Experiments were conducted on the above datasets to evaluate the performance of the proposed network on the above datasets. Consequently, results obtained through testing of multiple views has been analysed with respect to the proposed meta metric loss function against different metric losses to gauge the advantage of the proposed network. Further, different mining models from sec 3.3 were applied to test the efficacy of the meta metric learning for multi view sign language recognition. Finally, MVDMML is evaluated against other multi view models for robustness, ease of implementation and computational flexibilities.

4.1 KL_MV2DSL and multi view benchmark action datasets

The multi view sign language video dataset, KL_MV2DSL is captured with 5 cameras predominantly placed in the front view of the signer. The DSLR 16MP camera system has one camera focused on the signer in the front. Each of the two remaining cameras are oriented towards left and right of this centre camera with a horizontal angular displacement of 10 degrees between views. The sign language dataset is recorded with 5 test subjects in 5 different views. A total of 200 signs were selected from Indian sign language dictionary for creating multi view 2D video-based sign language, KL_MV2DSL. The entire MVSL dataset has 200×5×5=5000 sign videos. Each sign is recorded for 4 seconds at 30 fps. Currently, additional views and subjects are being appended to make our KL_MV2DSL a multi view sign language benchmark dataset.

Due to unavailability of multi view sign language datasets from other sources to evaluate the compatibility of the MVDMML, we worked with multi view human action datasets from NTU RGB D [13], MuHAVi [14], WEIZMANN [15] and NUMA [16]. The NTU RGB D is the largest dataset with 60 action classes in 80 views recorded with 40 subjects with a total sample size of 56880 videos of skeleton, depth and RGB. However, each action has been captured with 3 viewpoints at a time. We selected 60 action classes with 6 views from 40 subjects for training and testing the proposed loss function. The NTU RGB D dataset used in our work has 12000 video samples with 4 multi view representations of 40 subjects in 60 action classes.

The multicamera human action video data (MuHAVi) is recorded with 8 camera views placed rectangularly with 17 action classes performed by 14 subjects. We applied all eight views for training and testing MVDMML. The entire dataset was used for evaluation. However, WEIZMANN action dataset was most challenging one as it has only 90 videos of low resolution 180×144 with 9 subjects and 10 actions. The video sequences are recorded with 10 different viewpoints ranging from 0^0 to 81^0 in steps of 9^0 on only one side of the camera plane. During training we paired 0 to 36 degrees as left and 45 to 81 degrees as right views. The network size has been increased to 10 streams for operation on this dataset. Finally, northwestern-UCLA multiview action (NUMA) is a multi-view daily action dataset

Table 1. Comparison between state – of – the – art methods used on our sign language dataset KL_MV2DSL

Methods	mRA	mf1	m-P	m-f1	M-P	M-f1	r-1	r-5
VGG-16	0.6391	0.8994	0.8258	0.8151	0.7785	0.7475	0.8258	0.9311
GoogleNet	0.6441	0.8998	0.8758	0.8358	0.7859	0.7658	0.8825	0.9401
Resnet50	0.6056	0.8458	0.7785	0.7498	0.6975	0.6758	0.7985	0.8791
Inception V3	0.6323	0.8685	0.8235	0.7563	0.6875	0.6425	0.8491	0.9191
MVDMML	0.6512	0.8995	0.8875	0.8896	0.8286	0.7997	0.8798	0.9675

Table 2. Comparison between state – of – the – art methods used on NTU RGB D action dataset

Methods	mRA	mf1	m-P	m-f1	M-P	M-f1	r-1	r-5
VGG-16	0.6189	0.8752	0.8025	0.8036	0.7675	0.7127	0.8021	0.9042
GoogleNet	0.6245	0.8832	0.8456	0.8125	0.7741	0.7536	0.8495	0.9245
Resnet50	0.6085	0.8356	0.7458	0.7236	0.6803	0.6458	0.7571	0.8852
Inception V3	0.6259	0.8602	0.8125	0.7478	0.6889	0.6236	0.8125	0.9294
MVDMML	0.6358	0.8802	0.8584	0.8473	0.7869	0.7253	0.8579	0.9585

$$D_{\hbar}\left(q_j^{c'_{new}}, M_V^{c_{new}}\right) = \begin{cases} arg\max_i\left(d\left(f\left(q_j^{c'_{new}}\right), \max_{\forall M_V}\left(f\left(s_i^{c_{new}}\right)\right)\right)\right)\forall c'_{new} = c_{new} \\ arg\min_i\left(d\left(f\left(q_j^{c'_{new}}\right), \max_{\forall M_V}\left(f\left(s_i^{c_{new}}\right)\right)\right)\right)\forall c'_{new} = c_{new} \end{cases} \quad (13)$$

with 10 classes and 10 subjects. Inappropriately, there are only three views available for training and testing. Hence, the MVDMML now takes 3 views for training the architecture which is tested on different subjects in same views. Only RGB videos from benchmark action datasets were used for training and testing.

4.2 Evaluating the MVDMML framework

We evaluated the proposed MVDMML on our multi view sign language dataset with 4 training as well as test views. Eventually, all the task specific networks were kept constant across datasets with six CNN layers with leaky Relu activation with three maximum pooling layers after each two convolutional layers. Additionally, we have two batch normalization layers after 2nd and 4th convolution layers. We applied L2 weight regularization during training. The video frame size is 256×256×3. Consequently, all the network embeddings were computed by training with the proposed metric learning function in Eq. (13). The objective of this experiment is to test the proposed meta metric learning adaptability across state-of-the-art networks such as VGG-16, Resnet-50, GoogleNet and Inception V3 on all the datasets. All the networks were incepted from GitHub repository and are trained from scratch with hyperparameter tuning across datasets. The testing is performed with only one video view per class. We calculated the eight parameters on the multiple test views and the results are averaged over the tested views. Tables 1 to 5 shows the results on our multi view sign language dataset and the benchmark action datasets, respectively.

Singularly, our MVDMML model has shown to perform better on sign language dataset KL_MV2DSL and other benchmark action datasets. The meta metric loss embedding also produced exceptionally good performance on all the state – of – the – art models. Except the ResNet50, all have responded well to meta metric loss functional during the training process. The loss failed to impact on ResNet50 due to the gradient vanishing in deep layers. However, GoogleNet and InceptionV3 have shown to embrace the depth as the layers also move laterally in both these architectures. Despite their good performance, we found that GoogleNet and InceptionV3 consumed exponential training times when compared to our proposed MVDMML and VGG-16. Next, we compare the advantage offered by multi view meta learning loss functional against various loss functions.

Received: December 16, 2022. Revised: February 1, 2023.

Table 3. Comparison between state – of – the – art methods used on benchmark multi view action dataset MuHAVi

Methods	mRA	mf1	m-P	m-f1	M-P	M-f1	r-1	r-5
VGG-16	0.6086	0.7436	0.7832	0.7397	0.7603	0.7295	0.8002	0.9223
GoogleNet	0.6421	0.7853	0.8285	0.7682	0.7923	0.7836	0.8499	0.9398
Resnet50	0.6192	0.7258	0.8197	0.7482	0.7708	0.7672	0.8183	0.9225
Inception V3	0.6473	0.7557	0.8289	0.7599	0.7806	0.7769	0.8397	0.9422
MVDMML	0.6524	0.7645	0.8352	0.7708	0.7994	0.7909	0.8576	0.9449

Table 4. Comparison between state – of – the – art methods used on action dataset WEIZMANN

Methods	mRA	mf1	m-P	m-f1	M-P	M-f1	r-1	r-5
VGG-16	0.6162	0.7589	0.7956	0.7592	0.7896	0.7411	0.8115	0.9305
GoogleNet	0.6596	0.7996	0.8398	0.7905	0.8194	0.7796	0.8452	0.9444
Resnet50	0.6099	0.7549	0.7845	0.7499	0.7755	0.7498	0.7998	0.9142
Inception V3	0.6458	0.7758	0.8194	0.7805	0.8094	0.7698	0.8305	0.9342
MVDMML	0.6599	0.7799	0.8473	0.7997	0.8201	0.7705	0.8497	0.9597

Table 5. Comparison between state – of – the – art methods used NUMA dataset

Methods	mRA	mf1	m-P	m-f1	M-P	M-f1	r-1	r-5
VGG-16	0.6023	0.7345	0.7653	0.7269	0.7312	0.7043	0.7865	0.9086
GoogleNet	0.6386	0.7569	0.7856	0.7489	0.7492	0.7195	0.7936	0.9158
Resnet50	0.6004	0.7236	0.7523	0.7103	0.7092	0.7023	0.7812	0.9011
Inception V3	0.6236	0.7498	0.7754	0.7308	0.7211	0.7099	0.7899	0.9058
MVDMML	0.6365	0.7536	0.7803	0.7498	0.7392	0.7125	0.7826	0.9099

4.3 Comparison between metric loss functions

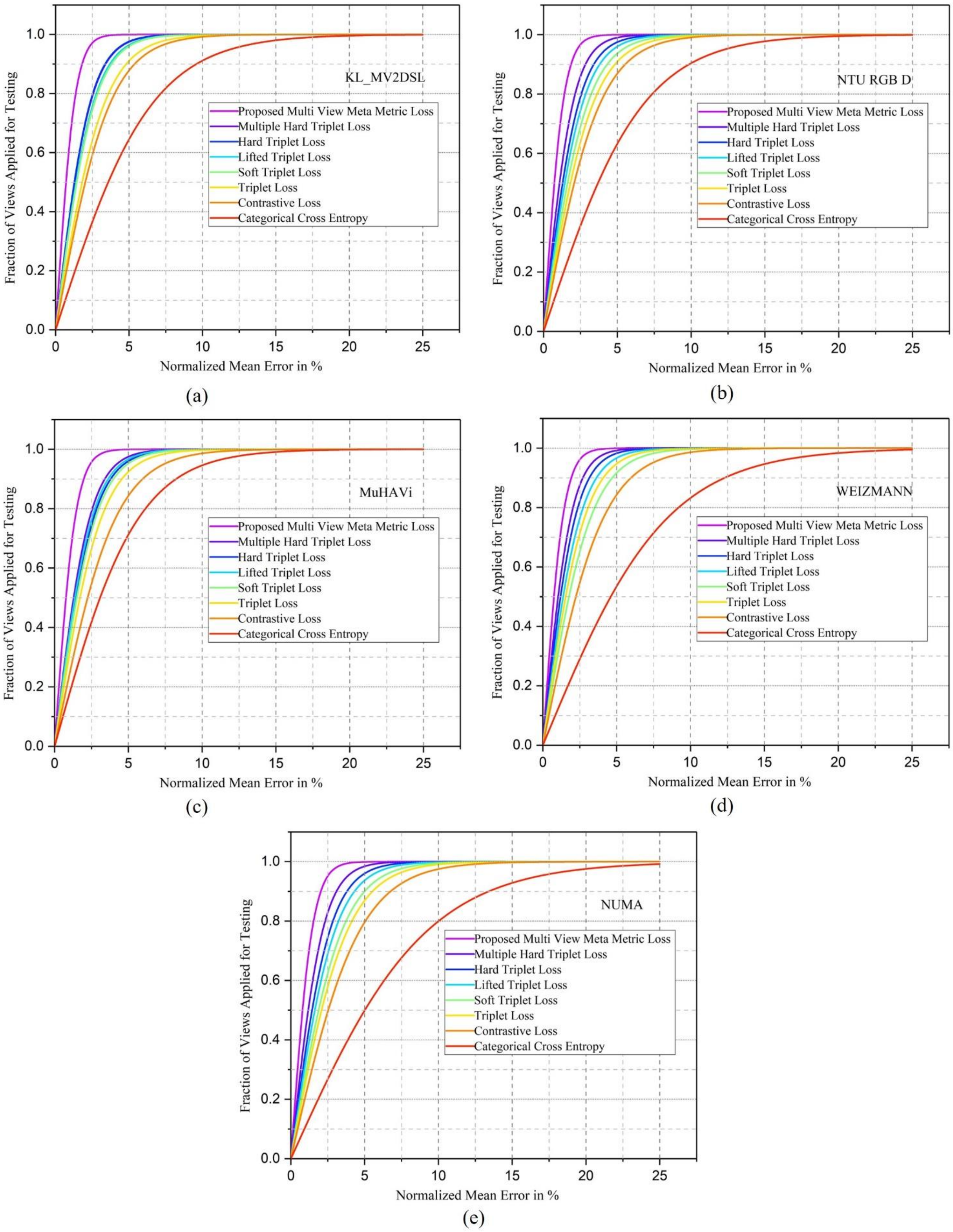
This section evaluates the performance of meta metric learning for multi view data based on the mean error obtained during testing with multiple view combinations. Accordingly, we conducted multiple experiments from single to available multiple views for testing the trained MVDMML on the considered sign (action) datasets. The results of the experimentation are presented as a plot between fraction of views applied for testing versus the normalized mean error. This experiment is performed on different kind of existing training loss functions to test the usefulness of meta metric loss proposed in this work. The following Fig. 5 shows the plots for our MVDMML model on our sign language dataset KL_MV2DSL and other benchmark action datasets. The mean error was computed as the difference between the predicted views and actual views. Earlier, the trained MVDMML is tested with different views on different loss functions. Some of the loss functions compared include multiple hard triplet loss [45], hard triplet loss [46], lifted triplet loss [47], soft triplet loss [48], triplet loss [49], contrastive loss [50] and categorical cross entropy [51].

The noticeable difference from the plots in Fig. 5 is that the metric loss performs better than categorical cross entropy due to the applications of paired samples in the former. Secondly, the triplet loss is better than contrastive loss which happens due to the selected margin between the positive and negative pairs.

4.4 Performance comparison with deep sign language methods

This section is dedicated to formal comparison of various sign language recognition models using video-based inputs. However, most of the methods in literature have focused on single view and we are assembling them along a few multi view sign language models. Table 6 shows the performance of sign language methods against our proposed method. The first 3 rows in Table 6 are multi view sign language methods which have similarity with the

Received: October 18, 2020. Revised: December 20, 2020.



(a)

(b)

(c)

(d)

(e)

Figure. 5 Cumulative multi view error distribution plots from MVDMML on: (a) KL_MV2DSL, (b) NTU RGB D, (c) MuHAVi, (d) WEIZMANN, and (e) NUMA datasets

Table 6. Comparison with deep sign language recognition methods

Reference	Methods	Datasets	Number of Views	% Accuracy
[25]	Elliptical Fourier descriptors and ANN	101 class Indian Sign words	4	82.27
[18]	Deep CNN with 8 layers	100 class Indian Sign Language words	1	90.56
[19]	Multi stream Deep CNN	200 Words from Indian Sign language	1	91.34
[21]	Multiple Deep Baseline Methods such as Recurrent CNNs, 3D CNNs, Graph CNNs.	2000 words of American Sign Language	1	84.63, 87.36, 92.56
[22]	BLSTM-3D residual networks	500 Daily Vocabulary words	1	86.9
[26]	CNN - BiLSTM	RWTH-PHOENIX-Weather multi-signer 2014 dataset	1	78.53
[27]	Iterative training using CNN (spatial) + CNN (Optical Flow) with feature summing.	RWTH-PHOENIX-Weather multi-signer 2014 dataset and SIGNUM signer-dependent se	1	85.63
MVDMML(Proposed)	Meta Metric Learning with Triplet Loss Embeddings	KL_MV2DSL dataset with 200 classes	5	96.75

proposed MVDMML. Otherwise, the rest are deep sign language models. The results show that the proposed meta metric embedding has outperformed the previously proposed methods. The high accuracy obtained for MVDMML is attributed to the models ability to discriminate overlapping features across views in different sign classes. Incidentally, we tested our model MVDMML on two sign language datasets in [21, 22]. Instead of using multiple views to form a meta cell, we used multiple subjects in these datasets to construct the meta cells. We used 500 classes from both the datasets for training and testing our proposed MVDMML. The average accuracy on dataset of [21] is 94.66% and that of [22] is 95.65% for a margin $\tau = 0.3$. This shows that the proposed MVDMML is better than the proposed networks in [21, 22] respectively.

It is observed that the increase in the number of support samples, the performance of the MVDMML has increased correspondingly. All the experiments were executed on NVDIA 8GB GTX1070. The results obtained show the effectiveness of using the proposed meta metric learning model MVDMML for multi view sign language recognition tasks.

5. Conclusions

In this work, we applied a meta metric learning model with set-based distances for multi view sign language recognition. In this model, we considered multiple views in randomly selected classes as meta cell and trained the deep model using samples from support and query sets in each episode. Consequently, the meta metric model learns by verifying the query sample with a margin-based cost function and maximum viewed pooled distance mining. The proposed MVDMML has been evaluated on our sign language (KL_MV2DSL) and other baseline action datasets. The meta metric model has shown improvement in multi view sign (action) language recognition tasks over state-of-the-art models.

Conflicts of interest

The author(s) declare that they have no Conflict of Interests for this research in any form.

Author Contributions

Substantial contributions to the conception or design of the work, ideas and theory formulation Venkata Vijay Kishore Polurie. Anil Kumar Dande contributed to the interpretation of the results. Suneetha Mopidevi and Prasad M.V.D., wrote the manuscript with input from all authors. All authors provided critical feedback and helped shape the research, analysis and manuscript.

References

[1] O. Koller, S. Zargaran, H. Ney, and R. Bowden, “Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs”, *International Journal of Computer Vision*, Vol. 126, No. 12, pp. 1311–1325, 2018.

[2] E. K. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. T. K. Kumar, and D. A. Kumar, “Training CNNs for 3-D Sign Language Recognition With Color Texture Coded Joint Angular Displacement Maps”, *IEEE Signal Processing Letters*, Vol. 25, No. 5, pp. 645–649, 2018.

[3] S. Ravi, S. Maloji, V. V. K. Polurie, and K. K. Eepuri, “Sign language recognition with multi feature fusion and ANN classifier”, *Turkish Journal of Electrical Engineering &amp; Computer Sciences*, Vol. 26, No. 6, pp. 2872–2886, 2018.

[4] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, “A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion”, *IEEE Sensors Journal*, Vol. 19, No. 16, pp. 7056–7063, 2019.

[5] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition”, In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] E. Hoffer and N. Ailon, “Deep Metric Learning Using Triplet Network”, In: *Proc. of International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92, 2015.

[7] J. Wang, K. C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, “Centroid-based Deep Metric Learning for Speaker Recognition”, In: *Proc. of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3652-3656, 2019.

[8] J. Yu, C. H. Hu, X. Y. Jing, and Y. J. Feng, “Deep metric learning with dynamic margin hard sampling loss for face verification”, *Signal, Image and Video Processing*, Vol. 14, No. 4, pp. 791–798, 2019.

[9] H. Coskun, D. J. Tan, S. Conjeti, N. Navab, and F. Tombari, “Human Motion Analysis with Deep Metric Learning”, *Lecture Notes in Computer Science*, pp. 693–710, 2018.

[10] J. He, Y. Wang, and H. Liu, “Ship Classification in Medium-Resolution SAR Images via Densely Connected Triplet CNNs Integrating Fisher Discrimination Regularized Metric Learning”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 4, pp. 3022–3039, 2021.

[11] N. Wojke and A. Bewley, “Deep Cosine Metric Learning for Person Re-identification”, In: *Proc. of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 748-756, 2018.

[12] G. Chen, T. Zhang, J. Lu, and J. Zhou, “Deep Meta Metric Learning”, In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9547-9556, 2019.

[13] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis”, In: *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010-1019, 2016.

[14] S. Singh, S. A. Velastin, and H. Ragheb, “MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods”, In: *Proc. of 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 48-55, 2010.

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time Shapes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 12, pp. 2247–2253, 2007.

[16] D. Wang, W. Ouyang, W. Li, and D. Xu, “Dividing and Aggregating Network for Multi-view Action Recognition”, In: *Proc of the European Conference on Computer Vision (ECCV)*, pp. 457–473, 2018.

[17] F. Pezzuoli, D. Corona, and M. L. Corradini, “Improvements in a Wearable Device for Sign Language Translation”, *Advances in Intelligent Systems and Computing*, pp. 70–81, 2019.

[18] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, “Deep convolutional neural networks for sign language recognition”, In: *Proc. of 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, pp. 194-197, 2018.

[19] S. Ravi, M. Suman, P. V. V. Kishore, K. E. Kumar, T. K. M. Kumar, and A. D. Kumar, “Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition”, *Journal of Computer Languages*, Vol. 52, pp. 88–102, 2019.

[20] P. V. V. Kishore, D. A. Kumar, A. S. C. S. Sastry, and E. K. Kumar, “Motionlets Matching

With Adaptive Kernels for 3-D Indian Sign Language Recognition", *IEEE Sensors Journal*, Vol. 18, No. 8, pp. 3327–3337, 2018.

[21] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison", In: *Proc. of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1459-1469, 2020.

[22] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic Sign Language Recognition Based on Video Sequence With BLSTM-3D Residual Networks", *IEEE Access*, Vol. 7, pp. 38044–38054, 2019.

[23] P. V. V. Kishore, D. A. Kumar, E. N. D. Goutham, and M. Manikanta, "Continuous sign language recognition from tracking and shape features using Fuzzy Inference Engine", In: *Proc. of 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2165-2170, 2016.

[24] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "Coupled HMM-based multi-sensor data fusion for sign language recognition", *Pattern Recognition Letters*, Vol. 86, pp. 1–8, 2017.

[25] P. V. V. Kishore, M. V. D. Prasad, C. R. Prasad, and R. Rahul, "4-Camera model for sign language recognition using elliptical fourier descriptors and ANN", In: *Proc. of 2015 International Conference on Signal Processing and Communication Engineering Systems*, pp. 34-38, Jan. 2015.

[26] R. Cui, H. Liu, and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization", In: *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7361-7369, 2017.

[27] R. Cui, H. Liu, and C. Zhang, "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training", *IEEE Transactions on Multimedia*, Vol. 21, No. 7, pp. 1880–1891, 2019.

[28] M. Kocabas, S. Karagoz, and E. Akbas, "Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1077-1086, Jun. 2019.

[29] Z. Gao, H. Z. Xuan, H. Zhang, S. Wan, and K. K. R. Choo, "Adaptive Fusion and Category-Level Dictionary Learning Model for Multiview Human Action Recognition", *IEEE Internet of Things Journal*, Vol. 6, No. 6, pp. 9280–9293, 2019.

[30] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 8, pp. 1963–1978, 2019.

[31] J. Zhu, W. Zou, Z. Zhu, L. Xu, and G. Huang, "Action Machine: Toward Person-Centric Action Recognition in Videos", *IEEE Signal Processing Letters*, Vol. 26, No. 11, pp. 1633–1637, 2019.

[32] Y. Zhu and G. Liu, "Fine-grained action recognition using multi-view attentions", *The Visual Computer*, Vol. 36, No. 9, pp. 1771–1781, 2019.

[33] K. Zhu, R. Wang, Q. Zhao, J. Cheng, and D. Tao, "A Cuboid CNN Model With an Attention Mechanism for Skeleton-Based Action Recognition", *IEEE Transactions on Multimedia*, Vol. 22, No. 11, pp. 2977–2989, 2020.

[34] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep Metric Learning with Angular Loss", In: *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2593-2601, 2017.

[35] X. Deng, W. Wu, and F. Wang, "Deep Metric Learning for text data based on Triplet Network", *IOP Conference Series: Materials Science and Engineering*, Vol. 806, No. 1, p. 012038, Apr. 2020.

[36] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep Metric Learning for Person Re-identification", In: *Proc. of 2014 22nd International Conference on Pattern Recognition*, pp. 34-39, 2014.

[37] J. Hu, J. Lu, and Y. P. Tan, "Discriminative Deep Metric Learning for Face Verification in the Wild", In: *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1875-1882, 2014.

[38] O. Ghahabi and J. Hernando, "Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 4, pp. 807–817, 2017.

[39] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification", In: *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1137-1145, 2015.

[40] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 5, pp. 2811–2821, 2018.

[41] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep Metric Learning with Hierarchical Triplet Loss", *Lecture Notes in Computer Science*, pp. 272–288, 2018.

[42] W. Zheng, Z. Chen, J. Lu, and J. Zhou, "Hardness-Aware Deep Metric Learning", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 72-81, 2019.

[43] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022-5030, 2019.

[44] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective", In *Advances in Neural Information Processing Systems*, pp. 1857-1865, 2016.

[45] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-Center Loss for Multi-view 3D Object Retrieval", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1945-1954, 2018.

[46] F. Qu, J. Liu, X. Liu, and L. Jiang, "A Multi-Fault Detection Method With Improved Triplet Loss Based on Hard Sample Mining", *IEEE Transactions on Sustainable Energy*, Vol. 12, No. 1, pp. 127–137, 2021.

[47] Z. He, C. Jung, Q. Fu, and Z. Zhang, "Deep feature embedding learning for person re-identification based on lifted structured loss", *Multimedia Tools and Applications*, Vol. 78, No. 5, pp. 5863–5880, 2018.

[48] M. Chen, Y. Ge, X. Feng, C. Xu, and D. Yang, "Person Re-Identification by Pose Invariant Deep Metric Learning With Improved Triplet Loss", *IEEE Access*, Vol. 6, pp. 68089–68095, 2018.

[49] X. Dong and J. Shen, "Triplet Loss in Siamese Network for Object Tracking", *Lecture Notes in Computer Science*, pp. 472–488, 2018.

[50] H. Choi, A. Som, and P. Turaga, "AMC-Loss: Angular Margin Contrastive Loss for Improved Explainability in Image Classification", In: *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[51] P. Zhong, D. Wang, and C. Miao, "An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss", In: *Proc of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 7492–7500, 2019.