Similarity Measure between Two Gestures using Triplets

Ravikiran Krishnan and Sudeep Sarkar
Department of Computer Science and Engineering
University of South Florida, Tampa, FL 33620, USA
{rkrishn2, sarkar}@cse.usf.edu

Abstract

One of the dominant approaches to gesture recognition, especially when we have one or few samples per class, is to compute the time-warped distance between the two sequences and perform nearest-neighbor classification. In this work, we show that we get much better results if instead we consider the similarity of the pattern of frame-wise distances of these two sequences with a third (anchor) sequence from the modelbase. We refer to these distance pattern vectors as the warp vectors. If these warp vectors are similar, then so are the sequences; if not, they are dissimilar. At the algorithmic core we have two dynamic time warping processes, one to compute the warp vectors with the anchor sequences and the other to compare these warp vectors. We select the anchor sequence to be the one that minimizes the overall distance, i.e. the sequence with respect to which these two sequences are the most similar.

We present results on a large dataset of 1500 RGBD sequences spanning 150 gesture classes, such as traffic signals, sign language, and every day actions, extracted from the ChaLearn Gesture Challenge dataset. We experimented with three different feature types: difference of frames, HOG and relational distributions. We found that there were improvements of 5%, 15%, and 7%, respectively, at 20% false alarm rate, over traditional two-sequence based time-warped distance.

1. Introduction

Human gestures are fast becoming the natural form of human computer interaction. This serves as a motivation to modeling, analyzing, and recognition of gestures. The large number of gesture categories such as sign language, traffic signals, everyday actions and also subtle cultural variations in gesture classes makes gesture recognition a challenging problem. Any gesture recognition task involves comparing an incoming or a query gesture against a training set of gestures. We call a collection of all the instances of all the classes available for training as a modelbase. These modelbases can have more than one instance per gesture class or they might have just one instance per class. If there is more than one instance then the recognition is based on learning statistics of features from the instances of the modelbase [7], [15], [5], [17], [19]. But this approach has its problems such as requiring large amounts of data to cover all variations of gesture classes or less of such leading to overfitting. There has been increasing interest in computer vision to avoid problems such as collecting and labeling large amounts of data, in a one-shot-learning approach for gesture recognition [1], [14], [6], [12]. While the term ''one-shot'' learning has been loosely used in the literature as one or few training instances, we consider it to refer to only one instance per class. We consider recognition in such a context. Methods given in [20], [2], [13], [8], all propose new

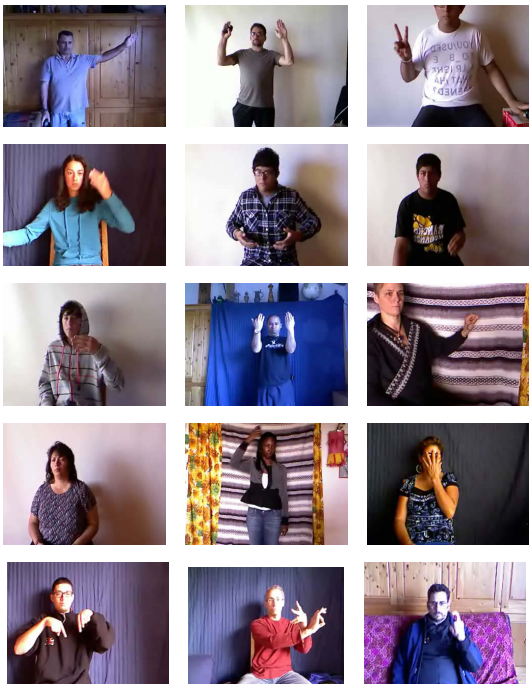


Figure 1: Examples of common gestures include sign language signals, traffic signals, hand signals and body language from ChaLearn Gesture Challenge Dataset [1].

one-shot similarity for images, but none of them use the one-shot-learning for gesture sequences.

Nearest-neighbor approach for gesture classification given a distance measure such as a time-warped distance is one of the dominant approaches in a one-shot-learning framework [11] [6]. In [6], performance on use of maximum correlation is experimentally shown. Based on [1], more than 50% of the proposed approaches uses time warped distance as a similarity measure in Chalearn gesture challenge. Recently, there have been approaches like [11] that uses Hidden Markov Models (HMM), where every frame is used as a state. We believe that this is similar to dynamic time warping in a probabilistic framework. Even though features used in computing similarity measure are important, we show that having a good similarity measure helps in boosting the performance of the classification. All of the measures proposed above consider the direct distance between an incoming query sequence to a model sequence. In this work, we show that we get better results if we included all training gestures when computing the similarity between a query and a model.

We introduce the notion of a third 'anchor' sequence to which we compute patterns (warp vectors) of frame-wise distances from the model and query sequences, respectively. The 'triplet' distance between these distance patterns is then obtained using a dynamic time warp process. We select the anchor sequence to be the one that minimizes the triplet distance, i.e, the sequence with respect to which model and query sequences are the most similar. In the process of selecting an anchor sequence, warp vectors from a model sequence to every other model sequence are computed which captures how varied a particular model is from every other model in the modelbase.

The idea of using such modelbase variations has been adopted as feedback on website text data [16]. This method uses the distance relation "A is closer to B than A is to C" to model data. Likewise, there are other relational clustering and boosting methods [10], [9], [21], [4]. None of the methods are used for one-shot-learning for gesture recognition.

The dataset used in our experiments are the gesture sequences extracted from the ChaLearn Gesture Challenge dataset [1]. Our dataset consists of 1500 sequences (both depth and RGB sequences) spanning 150 gesture classes. A depth sequence is a gray scale representation of the depth information. Each gesture in our dataset, has a depth sequence and a corresponding RGB sequence. Time-warping distance is used for baseline performance. We use three frame-based features in our experiments: difference of frames, Histogram of Oriented Gradients (HOG) [3], and Relational Distribution (RD) [18]. Each feature is applied to every frame in a particular gesture. HOG and relational distribution features are used on each frame of depth gesture sequences. The difference of frames is performed on the RGB sequence by subtracting the first frame with every following frame. We use the vectorized representation of these difference frames as our features. Figure 1 shows examples of different categories of gestures in the dataset. The notations used in this paper are summarized in Table 1.

Symbol	Meaning
$\mathbf{X_i}$	Model sequence.
$\mathbf{Q}$	Query sequence.
$\mathbf{f}_{\mathbf{X_i}}(k)$	Feature vector $\mathbf{f}$ corresponding to frame k in a particular sequence $\mathbf{X_i}$.
$s(\mathbf{X_i}, \mathbf{X_k} \mid \mathbf{X_j})$	s is a function that takes three gesture sequences as argument and returns a triplet distance between the sequence $\mathbf{X_i}$ and $\mathbf{X_k}$ conditioned on the third sequence $\mathbf{X_j}$(Sec 2).
$d(\mathbf{X_i}, \mathbf{Q})$	**Our Goal**: d is a function of s and returns the distance between the query sequence $\mathbf{Q}$ and model sequence $\mathbf{X_i}$.
$\mathbf{w}(\mathbf{u}, \mathbf{v})$	$\mathbf{w}$ is a **warp vector** of distances between corresponding frames (or volume of frames) in the sequence $\mathbf{u}$ and $\mathbf{v}$, where $\mathbf{u}$, $\mathbf{v}$ could be feature sequences or two different vector of distances $\mathbf{w_1}$ and $\mathbf{w_2}$. $\mathbf{w}$ is of recursive in nature.
$D(k, l)$	$\mathbf{D}$ is the distance matrix. Each entry is the Euclidean distance between the feature vector, $\mathbf{f}$ of frame k from sequence $\mathbf{X_i}$ to the feature vector, $\mathbf{f}$ of frame l from sequence $\mathbf{X_j}$.

Table 1: Summary of notations and terminologies used in this paper.

The layout of this paper is as follows: Section 2 shows how to calculate the triplet distance based on two warp vectors. Section 3 explains the use of the triplet distance to determine the anchor sequence and compute a new similarity (Refer to Table 1). Section 4 gives details about the experiments performed and the results achieved.

2. Triplet distance

Triplet distance is the concept of finding distance between two gesture sequences using a third (anchor) sequence. We consider the pattern (warp vectors) of frame-wise distances of two sequences with an anchor sequence. If these warp vectors are the similar, then so are the sequences; if not, they are dissimilar. Triplet distances are the fundamental blocks of the similarity measure proposed in the section 3. At the core of this distance, we have two time-warp processes, once to capture warp vectors and the other to compute the triplet distance between the warp vectors.

For explanation purposes, we assume that there are only two models sequences, $\mathbf{X_i}$ and $\mathbf{X_j}$ in the modelbase and a

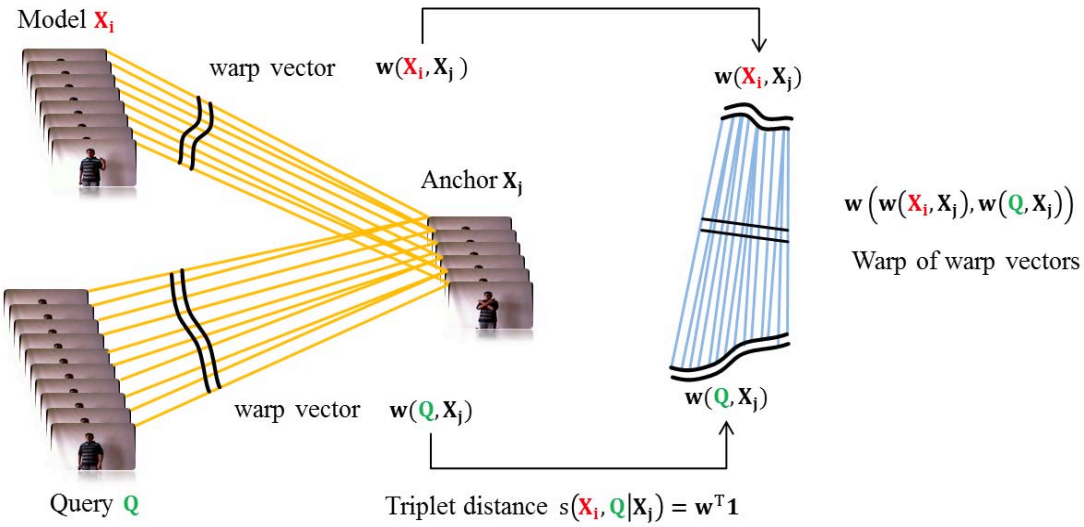


Figure 2: Conceptual illustration of triplet distance between three sequences: model sequence $\mathbf{X_i}$, anchor sequence $\mathbf{X_j}$ and a query sequence $\mathbf{Q}$. Warp vectors (time-warp path $\mathbf{w}$) between model sequence $\mathbf{X_i}$ and anchor sequence$\mathbf{X_j}$, and between query sequence $\mathbf{Q}$ and anchor sequence$\mathbf{X_j}$ are extracted. Dynamic time warp is applied between the two warp vectors $\mathbf{w}(\mathbf{X_i}, \mathbf{X_j})$ and $\mathbf{w}(\mathbf{Q}, \mathbf{X_j})$ to yield a distance between query sequence $\mathbf{Q}$ and model sequence $\mathbf{X_i}$. This triplet distance finds the minimized cumulative sum between the warp vectors.

query $\mathbf{Q}$. Our main goal here is to calculate how similar a model $\mathbf{X_i}$ is to query $\mathbf{Q}$ conditioned on another model $\mathbf{X_j}$. Warp vector, $\mathbf{w}$, captures the weights of the correspondences between frames based on their distances given by the non-symmetric matrix $\mathbf{D}$. This matrix is non-symmetric because each entry in this matrix is a distance from a frame in one sequence to a frame in another sequence and both sequences are allowed to have different number of frames. As the number the frames of query and model sequences can be of different lengths, we allow warp vectors to be also of varying length.

A detailed illustration of the triplet distance function, $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j})$ is given in Figure 2. In this figure, we have a gesture represented as set of images, warp vectors ($\mathbf{w}$) are represented as curves. In order to get the triplet distance, dynamic time warp is applied one more time on the two warp vectors to obtain $\mathbf{w}$ and a minimized distance.

In order to overcome varying length, a cost matrix between the two warp vectors is build that needs to be compared. Equation 1, takes the cost between the warp vectors as Euclidean and dynamic time warping process is performed once more on this cost matrix between the two warp vectors. Here also the time-warp path ($\mathbf{w}$) is a vector of distances between the two warp vectors and the sum of these distances gives us the triplet distance. This also is essentially comparing two distance matrices of different sizes. Dynamic time warp helps to maintain the time linear property of the gesture sequences. The value of $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j})$is greater than or equal to zero.

$$s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j}) = \mathbf{w}^{\mathbf{T}}(\mathbf{w}(\mathbf{X_i}, \mathbf{X_j}), \mathbf{w}(\mathbf{X_j}, \mathbf{Q}))\mathbf{1} \quad (1)$$

where $(\mathbf{w}(\mathbf{X_i}, \mathbf{X_j}), \mathbf{w}(\mathbf{X_j}, \mathbf{Q}))$ are the warped distance vectors obtained by performing dynamic time warp. The time-warp captures the distances which minimizes the distance between the pairs $(\mathbf{X_i}, \mathbf{X_j})$ and $(\mathbf{X_j}, \mathbf{Q})$. The vector of ones ($\mathbf{1}$) denote that all the values in $\mathbf{w}$ are summed together.

2.1. Warp Vector

Given a pair of gesture sequences, we want to capture a pattern vector from the distance matrix ($\mathbf{D}$) that has the frame-wise distances. We use dynamic time warping process and its resultant distances along the warp path to capture pattern vector and we call this as warp vector $\mathbf{w}$ between two gesture sequences. Before performing the time-warp process, we propose some pre-processing steps on $\mathbf{D}$, in order to speed up warp vector computation and noise reduction in distances. These two pre-processing goals are attained by averaging the distances in $\mathbf{D}$ over a temporal window R. We take the average in order to capture only those distances which capture the largest distance between a pair of frames. Similarly, the small changes will have smaller distance values. Hence, taking the average captures the larger changes as larger distances weigh more in each block. We call this process as noise reduction. We then normalize the averaged values and this pre-processing process is captured by the following equation:

$$D(k,l) = 1 - e^{\left(-\sum_{k-R}^{k+R}\sum_{l-R}^{l+R}||\mathbf{f}_{\mathbf{x_i}}(k) - \mathbf{f}_{\mathbf{x_j}}(l)||^2\right)} \quad (2)$$

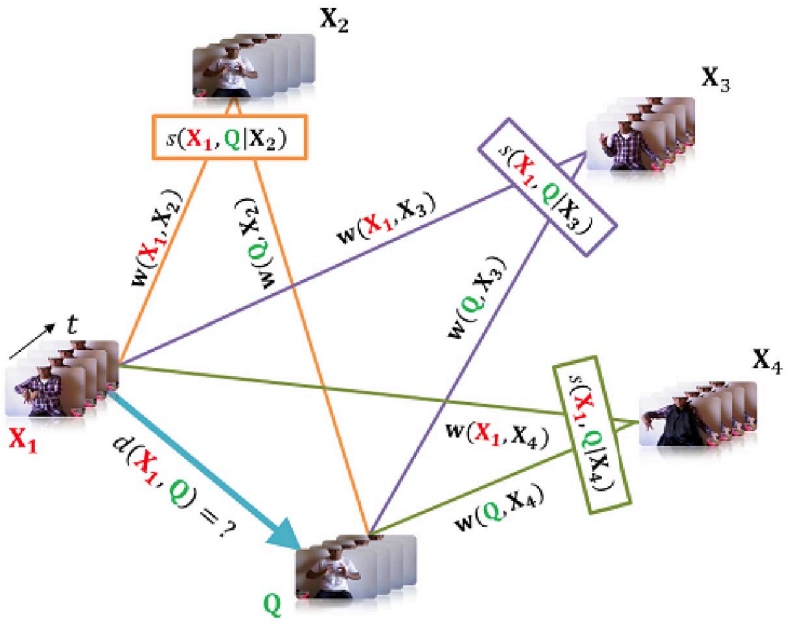


Figure 3: Our Goal: Distance $d(\mathbf{X_1}, \mathbf{Q})$ (directed edge) . Conceptual illustration of our proposed approach is shown here. There are 4 model sequences $\{\mathbf{X_1}, \ldots, \mathbf{X_4}\}$ and a query sequence $\mathbf{Q}$. The task is to compute a distance ($d(\mathbf{X_1}, \mathbf{Q})$) between model sequence $\mathbf{X_1}$ and query sequence $\mathbf{Q}$. The decision on $d(\mathbf{X_1}, \mathbf{Q})$ is based on a set of triplet distance s (Refer to Table 1 for notations). Model sequences ($\mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}$) are potential anchor sequences. Same color is used for the two undirected edges suggest that they both belong to the same triplet distance and capture the frame-wise distance pattern between the two connected sequences.

Where $l = \{1, \ldots, (K/R)\}$, $k = \{1, \ldots, (L/R)\}$. The first step towards building warp vectors, $\mathbf{w}(\mathbf{X_i}.\mathbf{X_j})$ is to extract features from gesture sequences $\mathbf{X_i}$ and $\mathbf{X_i}$. $\mathbf{f}_{\mathbf{X_i}}(k)$ is the feature vector $\mathbf{f}$ corresponding to frame k in the sequence $\mathbf{X_i}$. Any frame-wise feature can be applied, as long as the features capture motion and/or shape of the gesture. The Euclidean distance shown in Equation 2, gives the distance between a pair of frames. We have used three types of frame-wise features: difference of frames, histogram of gradient orientations (HOG) and relational distribution (RD). Distances in $\mathbf{D}$ are divided into equal size blocks $R \times R$. In each block, distances are averaged and these averaged distances are used to compute warp vectors.

3. Similarity using Triplet Distance

Let $\mathbf{M}$: $\{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_N}\}$, be the set of single instance model sequences. Each element $\mathbf{X_i}$ in $\mathbf{M}$ is a sequence that represents a particular gesture class. Our goal is to compute a distance $d(\mathbf{X_i}, \mathbf{Q})$ between query sequence $\mathbf{Q}$ and each model sequence $\mathbf{X_i}$. To take into account how a model sequence varies from other models in the model-base, we use the notion of triplet distances, $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j})$ (see Table 1 for notation details) based on the following idea. **If query sequence Q matches model sequence $\mathbf{X_i}$, then its distance to another model sequence $\mathbf{X_j}$, which we call an anchor sequence, should be similar to the variation between model sequences $\mathbf{X_i}$ and $\mathbf{X_j}$.** Triplet distance function $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j})$ is composed of two warp vectors, $\mathbf{w}(\mathbf{X_i}, \mathbf{X_j})$ and $\mathbf{w}(\mathbf{Q}, \mathbf{X_j})$, that define the relationship between sequences $\mathbf{X_i}$ and $\mathbf{X_j}$, and sequences $\mathbf{Q}$ and $\mathbf{X_j}$, respectively. $\mathbf{w}$, is defined as a vector of distances between each frame of one sequence aligned to every frame in another sequence. The triplet distance is a scalar value based on the comparison of the warp vectors $\mathbf{w}$. Lower the value, better the match between $\mathbf{Q}$ and $\mathbf{X_i}$. The distance $d(\mathbf{X_i}, \mathbf{Q})$ is then computed by taking the minimum of all triplet distances $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j})$ in the set M:

$$d(\mathbf{X_i}, \mathbf{Q}) = \min_{j \neq i} s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_j}) \tag{3}$$

This process is illustrated in Figure 3 using 4 model sequences and a query sequence $\mathbf{Q}$. In order to compute the similarity between a model sequence $\mathbf{X_1}$ and a query sequence $\mathbf{Q}$, we use triplet distances that are conditioned on model sequences $\mathbf{X_2}$, $\mathbf{X_3}$ and $\mathbf{X_4}$. Triplet distances are denoted by similarly colored edges that connect every triplet of videos in this figure. Each s is conditioned on a particular model sequence, which is a potential anchor sequence. The edges denote the pattern of frame-wise distances between two sequences. The directed edge denotes the new distance between $\mathbf{X_1}$ and $\mathbf{Q}$. This distance will always have an anchor sequence associated with it.

In order to give some insight into triplet distances, con-

sider this example, if $\mathbf{Q} = \mathbf{X_i}$, then $s(\mathbf{X_i}, \mathbf{Q}|\mathbf{X_1}) = 0$. This shows that the distances between the pair $(\mathbf{w}(\mathbf{Q}, \mathbf{X_1}), \mathbf{w}(\mathbf{X_1}, \mathbf{X_i}))$, are exactly the same. This would be the case for all the triplet distances in Equation 3. Hence, choosing the minimum out of these gives us the minimum distance between the query $\mathbf{Q}$ and model $\mathbf{X_i}$. And as query $\mathbf{Q}$ moves away from model $\mathbf{X_i}$, the distance between them increases. Moving away from another can be construed as $\mathbf{Q}$ moving closer to another class away from $\mathbf{X_i}$.

4. Discussion and Results

The dataset used in our experiments are the gesture sequences extracted from the ChaLearn Gesture Challenge dataset [1]. Our dataset consists of 1500 sequences (both depth and RGB sequences) spanning 150 gesture classes. The depth sequence is a gray scale representation of the depth information. The depth image has the advantage of being invariant to the appearance of the subject performing the gesture. For each gesture in our dataset, we have a depth sequence and a corresponding RGB sequence. The RGB and corresponding gesture sequences are divided into 15 batches. Each batch is a gesture of different category and every batch has 100 sequences out of which, 8 to 15 are model sequences and remaining are query sequences. The model gesture involves body language gesture, gestures which accompany speech, signs from sign language, traffic signals, every day actions such as drinking or writing, gestures made to mimic actions and dance postures. The query sequences contain 1 to 5 sequences in a single test instance. For the experiments in this paper, we use ground truth to manually segment query instances into individual gestures.

We use three different feature types and show performance over each of the feature type. We also show performances for each batch for one feature type and highlight the positive and negative gain in performance over the baseline performance. We show all visual results of gestures by representing them as motion history images. These images are for showing the movement and shapes of the gestures and were NOT used in our experiments.

Performance Measure: We evaluate the applicability of the distance measure and its effectiveness by computing performance as an ROC curve. This ROC curve represents a binary match, non-match test of the query sequences. We consider all the distances between the model and query sequences and test it against the ground truth. ROC curves are generated by varying a threshold variable. All the query sequences that were correctly matched, above a given threshold, are considered to be true positives. Similarly, all the query sequences that were incorrectly matched are considered to be false alarms. All the ROCs are shown up to 20% false alarm rate. Each ROC is built by averaging over all the batches in the dataset.

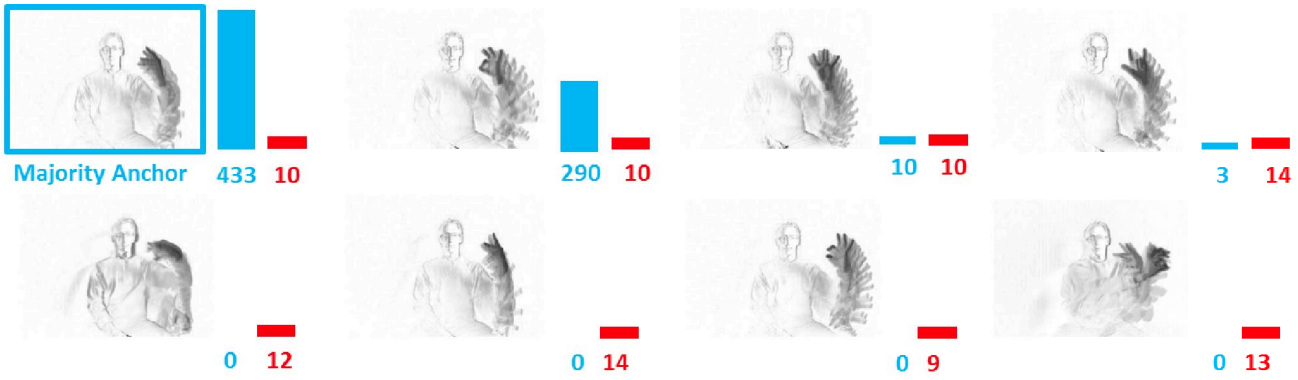
4.1. Anchor Sequence Analysis

Anchor sequence as explained earlier is the common element between two gesture sequences in the triplet distances. These anchor provide the relative information between query $\mathbf{Q}$ and model $\mathbf{X_i}$ sequences. When selecting the best similarity between two sequences, we look at the anchor that provides the distance minimizes the distance between the query and model sequence. In Figure 4, we show two sets of model sequences corresponding to two batches, each with 8 model sequences in its modelbase. We show gesture sequences represented as motion history images. This representation is for display purposes only, as it shows the motion and shapes involved in a particular gesture. The highlighted gesture is considered to be the majority anchor for a particular batch. This anchor was chosen 433 and 308 times in the modelbase of their respective batches. We cannot categorize a model as a majority anchor just by anchor selection count, as it depends on the number of times that particular model appears as a query sequence. There might be a case where a single model class could have the same number of anchor sequence selections. Hence, we have to look at the query distribution also and is important when labeling a model as the majority anchor. In Figure 4a and 4b, we can see that the anchor sequence that has the largest value does not equal the number of comparison of the query sequence with the highest instance count. The number of comparisons for this is $14 \times 8 = 112$, which much less than 433 and 308 anchor selection count. Although 1/3rd of the times the same anchor sequence is selected, making it the dominant anchor sequence for that modelbase, the determination of majority anchor is in this paper purely experimental.

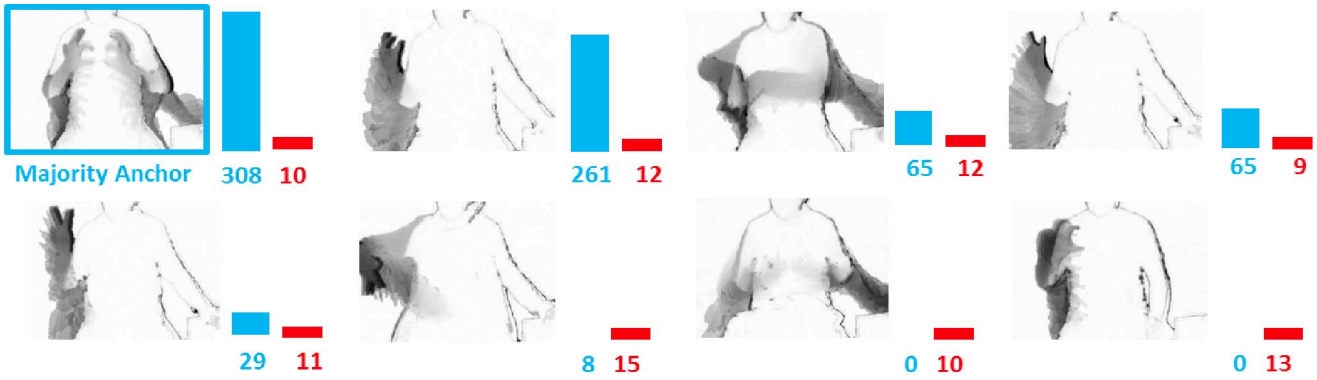
Min vs Mean: We compare the performance between two variants of choosing the anchor given by Equation 3. As similarity is a function of triplet distances, we consider minimization of all the triplet distances and the mean of the triplet distance as the two variants. Here we changed the minimization function into calculating the mean of all the triplet distances. This means that there is no minimized anchor video, but a mean of collection of all the anchor videos. The performance of these two variants are shown in Figure 6 for frame-wise HOG features. There is a dip in performance when mean of anchors were used. In another variant, instead of picking the minimum anchor sequence, we picked the maximum. This version performed poorly when compared to the baseline and hence not shown in Figure 6.

4.2. Performance

Our proposed method is compared with the baseline distance based on dynamic time warping process and performances are shown using ROC curves. Both the measures have the same set of features. Figure 7 (Left), shows the



(a)



(b)

Figure 4: Anchor sequence analysis for two batches. Model sequence ordered (left to right) based on the how many times (blue bar) a video was chosen as anchor sequence. Here the majority anchor (highlighted in blue) was chosen 433 times for batch in Figure 4a and 308 times for batch in Figure 4b. Both the majority anchors were represented as test in only 10 instances leading to $10 \times 8 = 80$ comparisons. The motion history images shown here are for representation purpose, **NOT** included in the calculation of anchor videos. Highlighted videos were the majority anchor.

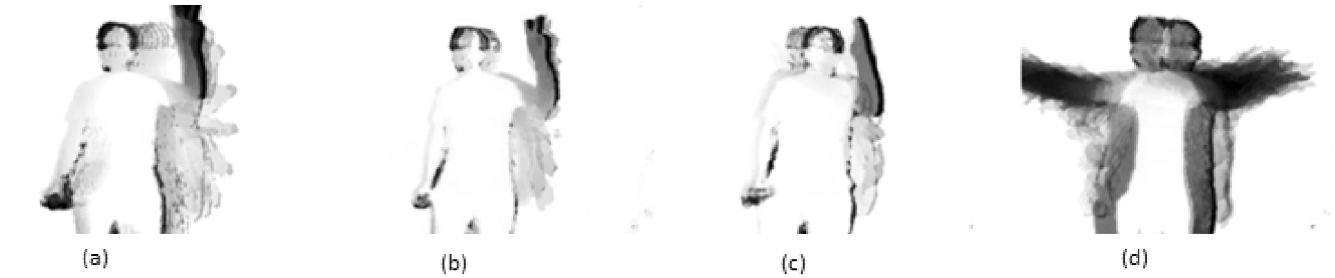


Figure 5: (a) Correctly matched sequence using proposed similarity measure, (b) Model sequence, (c) Mis-matched sequence using time-warp distance and (d) Chosen anchor sequence.

performance curves for HOG features. The dotted line in the plot below shows the method of using dynamic time warp with just the features and the solid lines represents our method. As shown in the performance of our method is better than the method using the features directly, with an improvement in detection rate of 5%. We have a constant window size (2 frames) for speedup and noise reduction. Match and non-match test based on proposed measure and time-warp distance is shown in Figure 5. The anchor sequence that was chosen for similarity is also shown and the same anchor is also a majority anchor.

We use two more feature types and generate ROC curves. Here the window size (2 frames) is again constant (chosen from a subset of classes) for the entire dataset and can be different for each feature type. In Figure 7 (Middle), we see the match, non-match results for the difference of frames

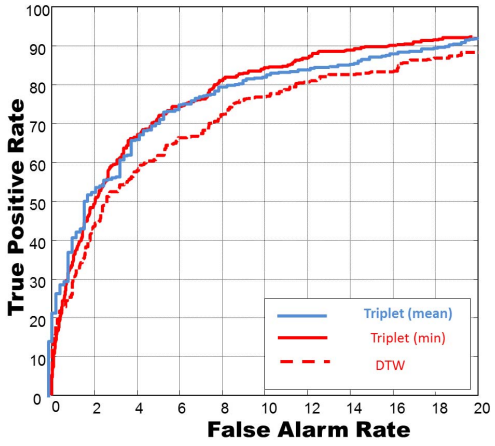


Figure 6: Comparing two different variants of our approach. The two variants are applied when minimum and mean of the anchor videos are considered. The ROC is plotted up to a false alarm rate of 20%.

(DOF) features and similarly Figure 7 (Right) shows the result for relational distribution features. The result does vary based on the features used. The performance between using dynamic time warp process based distance measure versus using the triplet distance based distance shows an improvement of around 15% and 7% for difference of frames and relational distribution as features respectively at 20% false alarm rate.

4.3. Batch Level Performance

We use 15 different batches and an example of each these batches are shown in Figure 1. These batches represent different categories of gestures performed. The results (Table 2) of our proposed approach when compared to picking the mean of the warp path from time-warp process vary in result depending on the category of gesture. All the performances shown in Table 2 are percentage of positive detections. The performances on six of the 15 batches have more than 90% accuracy in both the techniques, suggesting that the gesture category might not be challenging. Even if any increase in performance was possible, the increase as a percentage of positive detection would not be significant. The batches where there is performance degradation when using a particular feature type, our approach has shown to provide significant improvements. The improvement marks suggestions that features have successfully captured how varied modelbase is, and is used by our approach to decide on the similarity. The two batches that needs to be noted are Batch 14 and Batch 15 in Table 2. These two batches are of category – Surgeon Signals and Gang Hand Signals1. The positive detection rate depends on the number of query sequences in a particular batch. Each batch has a range of 85 to 92 query sequences. We use this fact as our measure to test whether the gain is significant or not. We take the error of removing one query from the each batch. As the batch have different number of query, we consider the number of query to be 100 and say that the error of removing one query is 1/100. If a gain is outside this error, then we say that the gain was significant. We do not consider gain for positive detection of more than 90% as significant. We have 7 batches out of 15 that has a gain percentage outside the error window, which makes them significant. The gain percentages are color coded in Table 2 – significant (green), non-significant (red).

5. Conclusion

In this paper, we have introduced the concept of triplet distances and warp vectors. We have shown the advantages of using warp vectors in conjunction with triplet distances in terms of improvement in performance one-shot learning framework. As we are choosing the anchor sequence from a set of model sequences that are not currently being compared, we can say that the similarity measure takes into account how varied a particular model is from every other model in the modelbase. This improvement in the result shows that the vector of distances should not be ignored. As the proposed approach captures how similar a particular gesture is from another gesture, this measure can be used in tasks such as clustering of gesture sequences. Even though the triplet distance was developed for frame-wise representation, we used the triplet distances for image representation and found that using the triplet distance approach did not give any improvement and more important is the fact that our approach did not hurt the performance over the direct distance between the query and model sequences. Similarly we can replace warp vectors computed from time-warp with other distance vectors that captures the frame-wise distance patterns. Our future work involves a theoretical analysis of this triplet distances and exploring the use of this similarity measure in other classification and clustering tasks.

References

[1] Chalearn gesture dataset (cgd2011-http://gesture.chalearn.org/2011-one-shot-learning. *ChaLearn, California*, 2011. 1, 2, 5

[2] One shot similarity metric learning for action recognition. In *Similarity-Based Pattern Recognition*, volume 7005, pages 31–45, 2011. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 2

[4] O. Danielsson, B. Rasolzadeh, and S. Carlsson. Gated classifiers: Boosting under high intra-class variation. pages 2673–2680, 2011. 2

[5] J. Davis and M. Shah. Recognizing hand gestures. In *In Proc. of European Conference on Computer Vision*, 1994. 1

[6] L. S. Di Wu, Fan Zhu. One shot learning gesture recognition from rgbd images. In *International Conference on Computer*

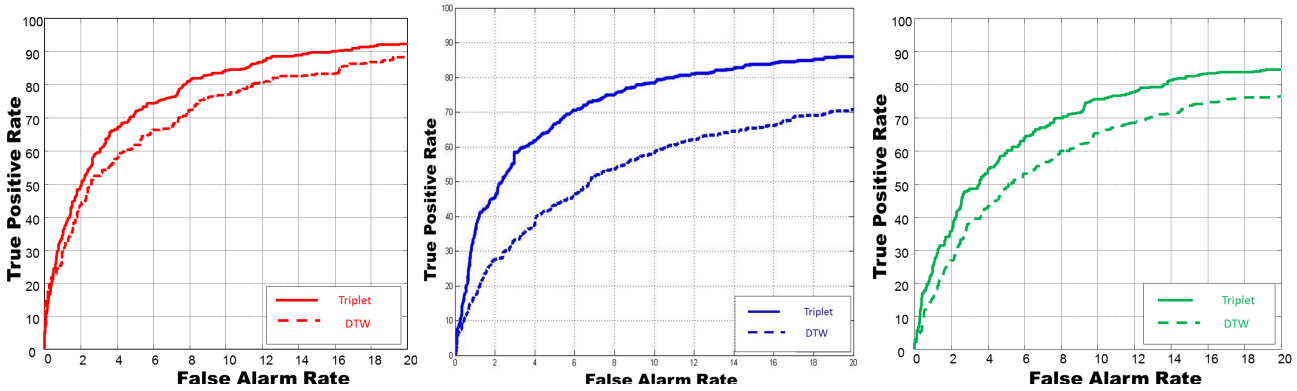


Figure 7: ROC curves for two matching methods our method (solid) and dynamic time warp (dotted). The three plots show the performances on different features - Histogram of Oriented Gradients (HOG) (Left), Difference of Frames (DOF) (Middle) and Relational Distribution (RD) (Right). The ROCs are plotted up to a false alarm rate of 20%.

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
DTW (%)	93.2±1	79.5±1	53.2±1	92.2±1	92.9±1	74.4±1	92.3±1	89.6±1	56±1	53.6±1	64.6±1	90.2±1	90.5±1	61±1	55.8±1
Triplet Approach (%)	92.8	85.2	64.3	91.2	96.7	86.6	96.6	93.4	70.3	74.6	80.3	90.4	87.3	91.1	84.3
Gain/Significance (%)	-0.4	5.6	11.1	-1.0	3.8	12.2	4.2	3.8	14.3	20.9	15.7	0.4	-3.2	**30.0**	**28.5**

Table 2: Table showing detection rates for each individual batch (consisting of 100 sequences) at fixed false alarm rate of 20%. The two gains from batches 14 and 15 are of gesture category surgeon signs and Gang Hand signals respectively, which yielded the top two improvements over baseline. The gain row also shows the significance test. The significant (green) results all show improvements over the baseline. The non-significant ones are marked in red.

Vision & Pattern Recognition, Workshop on Gesture Recognition, 2012. 1, 2

[7] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. pages 571–578, 2003. 1

[8] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006. 1

[9] R. J. Hathaway and J. C. Bezdek. Nerf *c*-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27:429–437, 1994. 2

[10] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, 22:205–212, 1989. 2

[11] E. Jackson. An hmm-based approach for gesture recognition using edge features. In *International Conference on Computer Vision & Pattern Recognition, Workshop on Gesture Recognition*, 2012. 2

[12] T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *British Machine Vision Conference*, 2004. 1

[13] O. Kliper-Gross, T. Hassner, and L. Wolf. In *SIMBAD, Lecture Notes in Computer Science*, pages 31–45. Springer. 1

[14] Y. M. Lui. A least squares regression framework on manifolds and its application to gesture recognition. In *International Conference on Computer Vision & Pattern Recognition, Workshop on Gesture Recognition*, 2012. 1

[15] S. Rajko, G. Qian, T. Ingalls, and J. James. Real-time gesture recognition with minimal training requirements and on-line learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, 2007. 1

[16] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2003. 2

[17] H.-I. Suk, B.-K. Sin, and S.-W. Lee. Hand gesture recognition based on dynamic bayesian network framework. *Pattern Recognition*, 43(9):3059–3072, 2010. 1

[18] I. R. Vega and S. Sarkar. Statistical motion model based on the change of feature relationships: human gait-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1323–1328, 2003. 2

[19] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115(2):224–241, 2011. 1

[20] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *IEEE International Conference on Computer Vision*, 2009. 1

[21] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, pages 505–512, 2002. 2