A Cause and Effect Analysis of Motion Trajectories for Modeling Actions

Sanath Narayan Kalpathi R. Ramakrishnan
Dept. of Electrical Engg., Indian Institute of Science, Bangalore
{sanath,krr}@ee.iisc.ernet.in

Abstract

An action is typically composed of different parts of the object moving in particular sequences. The presence of different motions (represented as a 1D histogram) has been used in the traditional bag-of-words (BoW) approach for recognizing actions. However the interactions among the motions also form a crucial part of an action. Different object-parts have varying degrees of interactions with the other parts during an action cycle. It is these interactions we want to quantify in order to bring in additional information about the actions. In this paper we propose a causality based approach for quantifying the interactions to aid action classification. Granger causality is used to compute the cause and effect relationships for pairs of motion trajectories of a video. A 2D histogram descriptor for the video is constructed using these pairwise measures. Our proposed method of obtaining pairwise measures for videos is also applicable for large datasets. We have conducted experiments on challenging action recognition databases such as HMDB51 and UCF50 and shown that our causality descriptor helps in encoding additional information regarding the actions and performs on par with the state-of-the art approaches. Due to the complementary nature, a further increase in performance can be observed by combining our approach with state-of-the-art approaches.

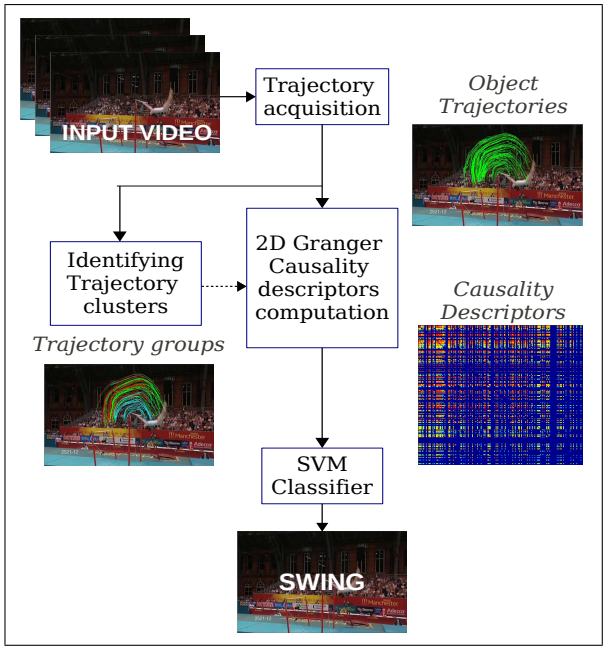

Figure 1. Illustration of the action classification approach: Object trajectories are obtained and clustered. The causal relations between the trajectories belonging to different clusters are computed and represented as a 2D histogram. These causality descriptors are compared against those in training and classified using a non-linear SVM classifier.

1. Introduction

Action recognition has been an important topic of research in the vision community for a long time. Human actions consist of space-time trajectories. Different actions have different trajectory patterns. Johansson [12] showed that humans can recognize actions by observing only few tracked joints. The motion trajectories belonging to an action are closely related to each other in terms of causality. Hence the trajectory patterns and their causal interdependencies vary from action to action. In this paper we exploit not just the occurrence of trajectory patterns but also the causal interactions among them. These interactions are directional. The interactions among different parts involved in an action are analogous to the interactions among different objects involved in a group-activity. The interactions give a richer description for the action classes. While the occurrence of features has been predominantly studied and analyzed in the present literature, the same cannot be said regarding the interaction among features. In this work, we focus on the interactions between trajectory pairs occurring in an action and model them using Granger causality measures.

1.1. Related Work

There have been many methods to classify actions based on space-time interest points (STIP) features using various

detectors based on Harris3D [14], separable Gabor filters [4], etc. Often local features for the interest points are based on gradient information, optical flow [4, 15, 22, 27], local trinary patterns [31], 3D-SIFT [23]. Other approaches for action recognition include space-time shape representations [7], template-based methods [2, 5, 20, 21].

Similar to the proposed work, few of the trajectory-based methods to perform action classification are presented in [1, 17, 29, 25, 18, 11]. Ali *et al*. [1] manually obtained the trajectories and used chaotic invariants as features to recognize actions. Matikainen *et al*. [17] used sparse trajectories from KLT tracker with elements of affine matrices in bag-of-words context as features. Messing *et al*. [18] track Harris3D interest points and use temporal velocity histories of trajectories as features. However, the performance of dense trajectories is seen to be better than sparse trajectories [29, 25].

Wu *et al*. [29] used chaotic invariants on dense trajectories to classify actions after segmenting motion through Robust PCA (RPCA) subspace learning technique [16].

Wang *et al*. [25] use local 3D volume descriptors based on motion boundary histograms (MBH) [3], histogram of oriented gradients (HOG) and histogram of optical flow (HOF) around dense trajectories to encode action. The MBH descriptors are known to be robust to camera motion. Recently in [26], Wang *et al*. estimate the camera motion and compensate for it and thereby improving the trajectories and the associated descriptors.

Recently, Jain *et al*. [10] decomposed the visual motion into dominant and residual motions for extracting trajectories and computing descriptors.

All the above approaches perform classification based on the presence of features or descriptors. Recently an action classification method using 2D histograms on trajectory features was proposed by Jiang *et al*. [11]. The relative position, motion direction and magnitude of relative motion between trajectory pairs (clustered based on different features viz. HOG, HOF and MBH) were used to construct the histograms. This approach involves computing feature interactions or dependencies at a coarse level since it uses low-level information of the trajectories. Our work computes the feature interactions based on causality between pairs of trajectories which is a mid-level information.

Causality has been used in other fields of research such as economics, biology [9, 6] to investigate directional relationships between signals. In computer vision it has been used to find the dominant flow information in video by Yamashita *et al*. [30], and for classifying different types of pair-activities like chasing, meeting, moving together, *etc*., by Zhou *et al*. [33]. While causality has been used to recognize actions previously [32], it has been applicable only in motion capture (MoCap) datasets. These methods had only few time series signals to deal with and the signals were known to be originating from specific joints in case of action recognition and specific neuron channels while investigating brain functions. Such a channel/body-joint based acquisition of motion trajectories cannot be expected in real-world videos. We propose an approach to incorporate the advantages of causality measures in applications where such constraints cannot be fulfilled.

Constructing the huge 2D descriptors (order of 10^6 to 10^7 entries in the histogram) for each video is time consuming and not viable from a practical standpoint. We provide a solution for constructing such descriptors quickly without sacrificing the relative structure among the descriptors.

Based on the above considerations, the *contributions of this paper* are:

i. An approach for incorporating the interactions among large number of signals (*i.e*., causal relations among motion trajectories in this work) in real-world videos has been proposed.

ii. A fast method for constructing large 2D descriptors providing a speed-up in computation time from $O(n^2)$ to approximately constant time has been proposed.

Organization of the rest of the paper is as follows. Approach for action recognition using interactions among features is explained in Section 2. The details of Experimental setup are provided in Section 3. Results on various datasets for action recognition and experiments related to different parameter settings of the causality descriptor are given in Section 4 and we conclude the paper in Section 5.

Figure 1 illustrates the proposed action recognition approach. The motion trajectories are acquired and their cluster memberships are identified. The causal interactions, *i.e*., Granger causality features between pairs of trajectories are computed resulting in a 2D histogram descriptor for the video. The action in the video is classified using an SVM classifier.

2. Interaction descriptors

Causal interactions occur inherently among motion patterns in an action and capturing such dependencies will help in modeling actions better. In this section we will formally introduce causality and use it to construct descriptors which will encode the dependencies. The computation of these descriptors for each video is time consuming. Hence we provide for an approach that will enable us to compute the measures once for an entire dataset and use them to construct individual descriptors for videos in the dataset.

2.1. What is Causality?

A signal is said to be "Granger causal" to another if the ability to predict the second signal is improved by incorporating the information about the first. This was formalized by Granger [8]. Given two time-varying signals $\mathbf{p}_t$ and $\mathbf{q}_t$,

if the error of predicting $\mathbf{p}_t$ using only the past samples of $\mathbf{p}_t$ is greater than the prediction error using past samples of $\mathbf{p}_t$ and $\mathbf{q}_t$, then $\mathbf{q}_t$ is said to Granger-cause $\mathbf{p}_t$. There are no restrictions on the complexity of the predictor functions. In this work we restrict the functions to be linear. Let $\mathbf{p}_t$, $\mathbf{q}_t \in \mathbb{R}^n$. Equation 1 and equation 2 are used to predict $\mathbf{p}_t$ using only the past samples of $\mathbf{p}_t$ and using past samples of both signals respectively.

$$\mathbf{p}_t = \mathbf{A}^T \mathbf{p}_{t-k}^{(m)} + \epsilon_1 \tag{1}$$

$$\mathbf{p}_t = \mathbf{B}^T \mathbf{p}_{t-k}^{(m)} + \mathbf{C}^T \mathbf{q}_{t-k}^{(m)} + \epsilon_2 \tag{2}$$

where prediction coefficient matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{mn \times n}$ and $\mathbf{p}_{t-k}^{(m)}$, $\mathbf{q}_{t-k}^{(m)}$ are k-delayed signals of order m, *i.e.*,

$$\mathbf{p}_{t-k}^{(m)} = (\mathbf{p}_{t-k}^T, \mathbf{p}_{t-k-1}^T, \ldots, \mathbf{p}_{t-k-m+1}^T)^T$$
$$\mathbf{q}_{t-k}^{(m)} = (\mathbf{q}_{t-k}^T, \mathbf{q}_{t-k-1}^T, \ldots, \mathbf{q}_{t-k-m+1}^T)^T$$

ϵ_1 and ϵ_2 are prediction errors and are assumed to be zero mean Gaussian noise with covariances Σ_1 and Σ_2 respectively. The causality ratio measuring the relative strength of causality from q to p is defined as

$$CR_{q \to p} = trace(\Sigma_1)/trace(\Sigma_2) \tag{3}$$

Causality ratio is an asymmetric measure. A high measure from q to p need not necessarily result in a strong measure from p to q. Since the error of predicting $\mathbf{p}_t$ can only decrease when incorporating additional information from the past samples of $\mathbf{q}_t$, the numerator of Equation 3 is always greater than the denominator. Hence the causality ratio is always greater than 1. Figure 2 illustrates the concept of Granger causality. The 1^{st} signal is following the 2^{nd}. The additional information from the 1^{st} signal while predicting 2^{nd} signal does not improve the prediction to the extent the additional information from the 2^{nd} signal while predicting 1^{st} signal does. Hence, the ratio $CR_{2 \to 1}$ is higher than $CR_{1 \to 2}$. With this reasoning, two coupled signals will have strong ratios in forward and feedback direction while two independent signals will have both causality ratios tending to 1 (indicating no causation).

2.2. Causality in actions

Actions consist of many space-time signals/trajectories. The trajectories are strongly inter-related and occur sequentially with a dependence on other motions which are part of the action. For walking action, the to and fro swing of one arm follows the to and fro swing of the other arm and thus, are dependent in terms of causality. This results in a strong causal measure. For a jumping action, both the knees move in the same way at the same time and incorporating one knee's motion does not improve the prediction of the other knee's motion. Hence the causal measures will not be

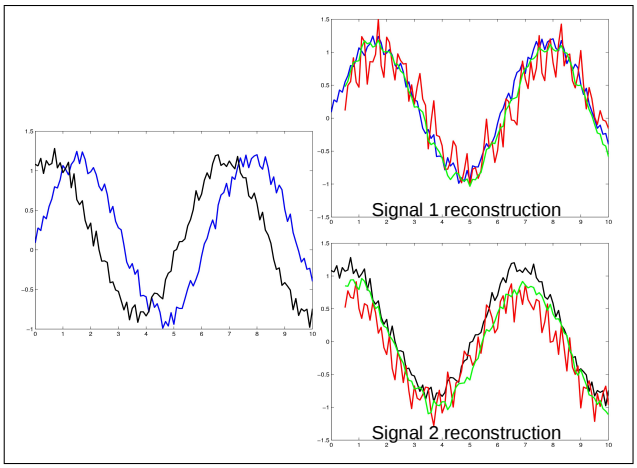

Figure 2. Illustration of Granger causality. Two 1D signals are plotted on the left. The 1^{st} signal (blue) is following the 2^{nd} (black). On the right, each plot shows the original signals, predicted signal (in red) using only its own previous samples and another predicted signal (in green) using previous samples of both signals. Prediction improves for both signals when information from other signal is used. However, the improvement in prediction for 1^{st} signal when previous samples of 2^{nd} signal are also used is higher compared to the improvement for 2^{nd} signal when previous samples of 1^{st} signal are used.

strong in this action. Figure 3 shows different actions having different kinds of interactions. The strength of causality from one node (representing a body part) to another is differentiated by thickness and color of the corresponding edge. For "jack" action, the strength of causality is high between the two arms. For walking action, there is a moderate Granger causation from legs to arms and a weak one from arms to head. It should be noted that, Granger causality represents "predictive causality". It may or may not represent the "true causality".

Causality ratios are between pairs of trajectories and number of trajectories acquired from actions in real-world videos vary from one video to another. Since causality matrix will be of size of squared number of trajectories, a direct comparison becomes difficult. Hence, the trajectories are clustered in a feature space into K words. The cluster centers are obtained initially by clustering a subset of trajectories of the training videos. The causality ratios of pairs of trajectories are used to obtain the ratios for pairs of clusters. A causality matrix of $K \times K$ is used to represent the ratios between clusters. Element (i, j) of the matrix is the average of causality ratios of trajectories of i^{th} cluster that are "Granger-caused" by the trajectories of j^{th} cluster. The matrix is normalized and represents a 2D normalized histogram which encodes the interactions between trajectory clusters. Thus, each video is now represented by a causality descriptor of size $K \times K$.

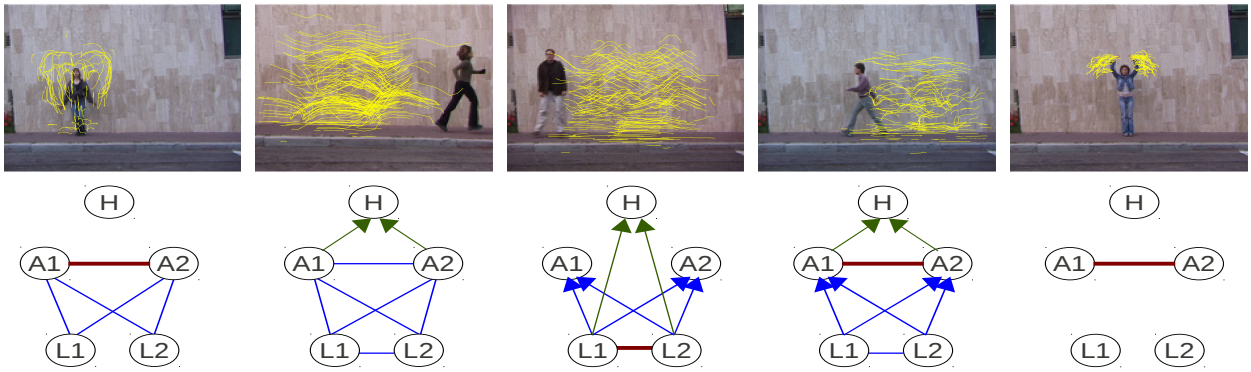

Figure 3. Figure shows five different actions from Weizmann dataset [7] and their corresponding causal graph structures. From left, the actions are Jack, Run, Side, Walk, and two hand Wave. Thick red lines represent strong causal relations. Medium thick blue lines represent medium strength and thin green lines represent low causal relationships. Absence of line between 2 nodes indicates causation is absent. Directed line indicates that the causality ratio is high in that direction compared to the opposite direction. Undirected line indicates that ratios are similar in both directions. The graphs are represented with nodes of body parts Head, Arm1, Arm2, Leg1, and Leg2. This illustrates that different actions have different causal graphs.

2.3. Overcoming computational bottleneck

The causality descriptor is constructed from causality ratios between pairs of trajectories. Since the ratios in the forward and feedback directions are asymmetrical, N trajectories will require nearly N^2 computations of causality ratio resulting in three prediction matrices to be learnt for every single ratio separately. Given that a video can typically have thousands of trajectories, it becomes practically infeasible to compute the causality descriptor.

One way to overcome this would be to generate the measures between pairs of clusters. We find the mean of the trajectories of every cluster and compute the pairwise ratios between the mean trajectories and consider them as pairwise cluster measures. Here, we compute ratios between the mean trajectory of one cluster and the mean trajectory of another cluster, unlike previously (Section 2.2) where it was the average of ratios from all pairs of trajectories belonging to two different clusters. This can be computed once for the entire dataset, when the clustering is done in the feature space. This results in a reference matrix, denoted by R, of causal measures of size $K \times K$ and requires K^2 causal ratio computations.

We use the reference matrix generated to obtain causal descriptors for individual videos. The number of trajectories belonging to each cluster is used as a scaling factor to compute the causal interactions between the clusters. The video causal descriptor, denoted by CD, is of same size $K \times K$. The (i, j) entry of the descriptor is the corresponding entry of the reference matrix scaled by the number of trajectories in i^{th} and j^{th} cluster.

$$CD(i, j) = N_i N_j R(i, j) \tag{4}$$

where N_i and N_i are the number of trajectories belonging to the i^{th} and j^{th} cluster respectively and $i, j \in [1, K]$. Since we are using the precomputed reference causality matrix values, computing causality ratios for trajectory pairs in each video is avoided by this approach and this saves a lot of computational time. The 2D descriptor is l_1-normalized resulting in a 2D histogram.

3. Experimental setup

In this section, the details of the experimental setup with various parameter settings are provided. We present a baseline descriptor for comparing the performances of action classification on different datasets. The details of constructing the baseline descriptor are given in section 3.3. The datasets that are used are for evaluating the approach are also presented in section 3.4.

3.1. Trajectory Acquisition

Using the "particle" concept [29, 28], a pixel in a video frame corresponds to a particle and the motions in the scene are represented quantitatively by the motions of the particles using dense optical flow. A particle $\mathbf{p}_t = (x_t, y_t)$ at frame $t \in [1, L-1]$ moves to

$$\mathbf{p}_{t+1} = (x_{t+1}, y_{t+1}) = \mathbf{p}_t + (u_t(\mathbf{p}_t), v_t(\mathbf{p}_t)) \tag{5}$$

at frame $t+1$. Here (u_t, v_t) represent the optical flow field at frame t. By advecting the particles in consecutive frames, a trajectory is obtained and represented as $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_L)$. We use the code [1] provided by [25] with default settings to acquire the trajectories. The length of the trajectories is 15 samples. We obtain the actual trajectories of the video

[1] http://lear.inrialpes.fr/people/wang/improved_trajectories

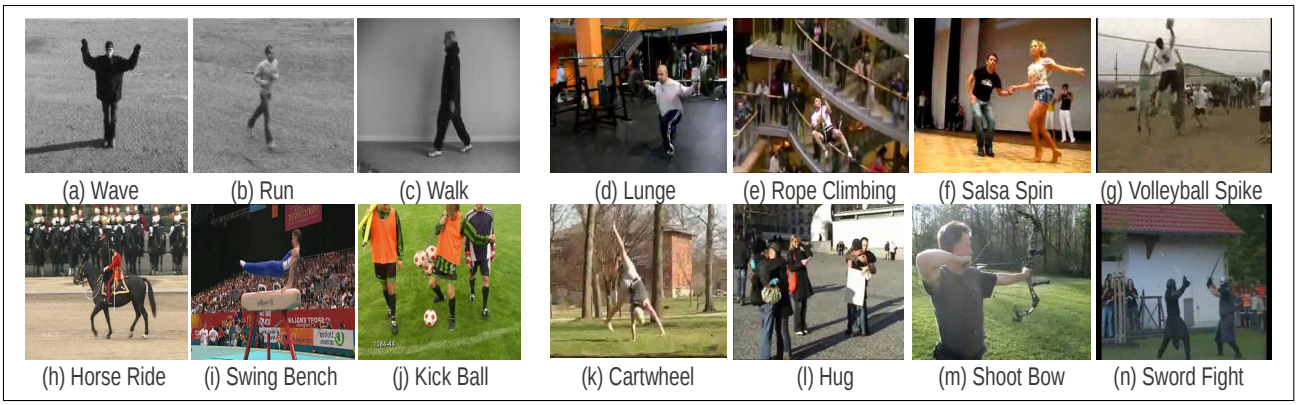

Figure 4. A few sample actions from the datasets are shown. (a)-(c) are actions from KTH, (d)-(g) are from UCF50, (h)-(j) depict actions from UCF Sports and (k)-(n) are from HMDB51.

from the code in addition to the default TrajShape feature as output by the code. The trajectories thus obtained are zero-mean normalized. Other features of trajectories, *viz.*, HOG, HOF and MBH are also obtained.

3.2. Causality Descriptor

The trajectories are clustered in 4 feature spaces, *viz.*, TrajShape, HOG, HOF, and MBH. The dictionary size is chosen to be 4000 words. $100,000$ feature samples are randomly selected from the dataset for k-means clustering. The clustering in MBH feature space is done separately for x and y spatial dimensions. We choose a 3^{rd} order model and a 3 sample delay (Equation 1) for computing causality measures. Standard regression function is used to estimate the predictor coefficients of equations 1 and 2. We experiment with different values for order and delay and report the results in Section 4.3.

Since causal measure from one motion to another is meaningful only within a time-frame for actions (*e.g.* within action cycles), it is required to restrict the process of populating the pairwise measures within a segment of the video. We consider non-overlapping 40-frame segments to compute the local descriptors of causality. The global causality descriptor is obtained by summing the local descriptors. In order to preserve the relative strength of causality among the pairs in the local descriptors, the local descriptors are l_1-normalized. The single global descriptor for the video is also l_1-normalized. The causality descriptor is of size 4000×4000 and it is vectorized to a 1.6×10^7-d vector for the purpose of implementation.

The similarity between two descriptors is computed using histogram intersection kernel. We use a *one-vs-all* approach while training the multi-class classifier. Different descriptor types are combined by soft-fusion method of averaging the corresponding kernel scores.

3.3. Baseline descriptor

To evaluate our 2D descriptor, we make a baseline comparison and show the effectiveness of our causality descriptor. The baseline descriptor is also a 2D descriptor and is computed using correlation coefficients between pairs of trajectories. Correlation coefficient, to an extent, can measure the interaction between trajectories pairs. However the direction of causation cannot be known through correlation measures.

The descriptor is constructed using the same steps the causality descriptor is constructed with and under the same settings such as cluster trajectories, dictionary size. However the values of correlation coefficients lie in the range $[-1, 1]$. This may result in an erroneous global descriptor when the local causality descriptors are summed. Hence the coefficients are transformed to the range $[0, 1]$ in the reference matrix before populating the local descriptors. Since correlation coefficient is symmetric, the global correlation descriptor is also symmetric matrix of 4000×4000. But there are only ${}^{4000}C_2$ independent entries in the descriptor. We compute the coefficients separately for x and y dimensions of the trajectories. Hence ${}^{4000}C_2$ independent entries of both descriptors are concatenated to form a single histogram of nearly 1.6×10^7 dimension. A non-linear SVM with histogram intersection kernel is used for the classification stage. We use a *one-vs-all* approach while training the multi-class classifier.

3.4. Datasets

We perform the experiments on four action recognition datasets and report the results. The datasets used for evaluating our work on interaction descriptors are KTH, UCF Sports, UCF50 and HMDB51. The UCF50 and HMDB51 datasets are large-scale databases containing over 6600 videos each. The datasets are from uncontrolled settings (except for KTH), with changes in viewpoints, il-

lumination, camera motion, background clutter and hence are very challenging. A few sample actions from the four datasets have been shown in Figure 4.

The **KTH dataset** [22] has close to 600 videos for 6 actions performed several times by 25 subjects with different settings. The action classes are walking, jogging, running, boxing, waving and clapping. The background is homogeneous and static in most sequences. There are 2391 video samples in the dataset. The dataset is split into training set (16 subjects) and testing set (remaining 9 subjects) as in the original setup [22].

The **UCF sports** dataset [20] contains 10 human action classes, *viz*., diving, golf swinging, kicking a ball, weight lifting, horse riding, running, skateboarding, swinging (bench), swinging (high bar) and walking. There are 150 videos in the dataset and a large intra-class variability is present in the videos. The samples are increased by adding a horizontally flipped version of the videos to the dataset. We use leave-one-out cross validation as in Wang *et al*. [27] where the flipped version of the test video is removed from the training set. Since the video clips are shot at 10 fps, only for this dataset, we consider the non-overlapping segments of 10 frames to compute the local causality and correlation descriptors.

The **UCF50** dataset [19] is an action recognition data set with 50 action categories, consisting of realistic videos taken from YouTube. The actions categories vary widely from general sports to playing musical instruments to daily life exercises. For all 50 classes, the videos are split into 25 groups. For each group, there are at least 4 action clips. In total, there are $6,618$ video clips. The video clips in the same group may share some common features, such as the same person, similar background or similar viewpoint. We apply the Leave-One-Group-Out Cross-Validation (25 cross-validations) as suggested by the authors [19].

The **HMDB51** action dataset [13] is collected from various sources, mostly from movies, and from public databases such as YouTube and Google videos. The dataset contains 6766 clips categorized into 51 action classes, each containing a minimum of 101 clips. The action categories can be grouped into general facial actions, general body movements with and without object interactions and human interactions. We use the original 3 train-test splits provided by the authors for evaluation. Each split contains 70 videos and 30 videos from every class for training for testing respectively. The average classification accuracy over the three splits is reported.

4. Experimental results

In this section, we present the results of our approach evaluated on the two datasets. We first discuss the performance of our approach in comparison with the baseline approach in section 4.1. In section 4.2, we compare the performance of causality descriptors with the state-of-the art techniques and in section 4.3, we investigate the effects of varying the order and delay parameters of causality.

With K set to 4000 words, it took approximately 10 minutes in MATLAB on an i7 (3.5GHz) machine without parallelization for computing the reference causality matrix R in equation 4. The computation time for the causality descriptor (including the 1D bag-of-words descriptor) for individual videos was 1.3 to 1.4 times the computation time required for the 1D descriptor alone.

4.1. Comparison with Baseline

In this section, we compare the performance of our causality descriptor with that of the baseline descriptor described in section 3.3 on the four datasets discussed in section 3.4. Both the descriptors encode interactions among the motions of the action being performed. The causality descriptor models the causation among the interactions while the baseline descriptor based on correlation does not. Correlation does not differentiate whether one motion trajectory is causing the motion of another trajectory or not. In both situations, it results in the same quantified measure. We observe from the experiments that the causality descriptor always performs better than the baseline descriptor. The reason for this is that the property of cause and effect is exploited in the causality descriptor. The asymmetrical nature of the causality measure helps in encoding the interactions among the motions better than the baseline correlation descriptor.

We have reported the performance of both descriptors in Table 1 for the four datasets. The causality descriptor outperforms the baseline on all of them. The correlation and causality descriptors are computed in 4 feature spaces. The performance of the descriptors in each feature space are reported. The combined performance of the descriptors in different feature spaces are also reported.

4.2. Comparison to state-of-the-art results

In this section, we compare the performance of our combined descriptor with the state-of-the-art results till date.

On KTH dataset, we perform on par with the other results in literature as shown in Table 2. A classification accuracy of 96.5% was achieved. The error was mainly due to misclassifying some samples of running as jogging. The samples were those which were difficult even for humans to always correctly classify.

Table 2 also gives the comparison on UCF Sports dataset. Our approach achieves an average accuracy of 92.8%. We perform better than the other trajectory based action recognition approaches. Only Sadanand and Corso [21] do better than our approach. However, they achieve 26.9% on HMDB51 and 76.4% (using a different cross-validation) on UCF50. We do better on these large datasets as seen below.

Approach	Feature Space	KTH	UCF Sports	UCF50	HMDB51
Baseline	TrajShape	90.1%	76.3%	71.4%	30.9%
	HOG	87.8%	84.5%	75.5%	36.1%
	HOF	93.3%	77.2%	82.8%	46.6%
	MBH	94.3%	84.1%	85.2%	49.0%
	Combined	95.3%	89.5%	87.5%	51.1%
Causality	TrajShape	90.6%	77.1%	73.8%	31.3%
	HOG	88.2%	85.9%	80.4%	38.2%
	HOF	94.1%	78.5%	85.3%	47.1%
	MBH	95.4%	88.7%	87.5%	50.8%
	Combined	**96.5%**	**92.8%**	**89.4%**	**53.4%**

Table 1. Performance comparison on different datasets using baseline descriptor and causality descriptor. The descriptors are computed in 4 feature spaces and performance for each is reported. The performance of the combination of descriptors is also tabulated for the baseline and causality approaches.

Method	KTH	UCF Sports
Sadanand & Corso [21]	98.2%	95.0%
Proposed	96.5%	92.8%
Wu *et al*. [29]	95.7%	89.7%
Wang *et al*. [25]	95.3%	89.1%

Table 2. Performance comparison on KTH and UCF Sports datasets

UCF50		HMDB51	
Wang* *et al*. [26]	91.2%	Wang* *et al*. [26]	57.2%
Wang$^+$ *et al*. [26]	87.2%	Wang$^+$ *et al*. [26]	52.1%
Shi *et al*. [24]	83.3%	Jain *et al*. [10]	52.1%
Reddy *et al*. [19]	76.9%	Jiang *et al*. [11]	40.7%
Proposed	89.4%	**Proposed**	53.4%
Combined	92.5%	**Combined**	58.7%

Table 3. Performance comparison on UCF50 and HMDB51 datasets. The * and $^+$ for Wang *et al*. represent Fisher and bag-of-words encodings respectively. 'Combined' represents the performance when we combine the Fisher encoding of [26] and our method. This illustrates the complementary nature of our descriptor.

The performance of our causality descriptor on the large datasets has been tabulated in Table 3. On UCF50, we achieve an accuracy of 89.4%. Improved trajectories approach by Wang *et al*. [26] achieves 87.2% with bag-of-features encoding and 91.2 with Fisher vector encoding. On HMDB51, the method achieved an accuracy of 53.4% while Wang *et al*. achieves 52.1% and 57.2% with bag-of-features and Fisher vector encodings respectively. Our method performs better than the bag-of-words encoding method of Wang *et al*. but slightly below par when compared to the Fisher vector encoding approach. Since our method is based on clustering feature spaces using bag-of-features, its performance is lesser than that of the Fisher encoding approach. However, due to complimentary nature of our descriptor, when combined with the Fisher encoding approach of Wang *et al*. [26], we observe an increase in the classification accuracy. The methods are combined by averaging the kernel similarity measures. The combination achieved 92.5% and 58.7% on UCF50 and HMDB51 respectively. Thus, our descriptor can be combined with other methods to improve the performance.

4.3. Varying the order and delay parameters

In this section, we investigate how the performance of our causality descriptor varies as the order and delay parameters (equation 1) are changed. Keeping the delay constant at 3, we vary the order of the model from 1 to 5 and plot the accuracy of classification as a function of model order for UCF50 and HMDB51 datasets. From figure 5, we observe that as model order increases, the accuracy also increases initially but decreases with further increase of the order parameter. Smaller values of order will not be very effective in capturing the dynamics of the trajectories. For large order values, the number of equations available for computing the prediction matrices decrease resulting in noisy values for causality ratio. Changing the delay parameter for smaller model orders did not affect the performance while for larger order values the performance degraded with increase in delay due to the decrease in number of equations for computing prediction matrices.

5. Conclusion

We have proposed novel descriptors, for action recognition, which are based on the causal interactions among the motion trajectories. These interactions play a vital role in the dynamics of actions. This work highlights the importance of obtaining pair-wise causal informations for action classification along with the individual occurrences of

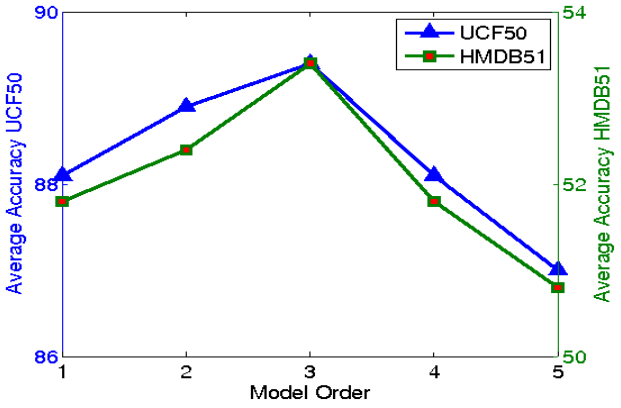

Figure 5. Plot showing the classification accuracy for UCF50 and HMDB51 for varying model orxder parameter

motion information. We have evaluated our proposed approach on challenging action recognition datasets and have shown that capturing interactions will help in modeling actions better and in providing additional information about the actions.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *TPAMI*, 23(3):257–267, 2001.

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior a recognition via sparse spatio-temporal feature. In *VS-PETS*, 2005.

[5] A. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[6] K. Friston. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7(2):e1000033, 2009.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29(12):2247–2253, 2007.

[8] C. W. J. Granger. Investigating causal relations by econometric models and cross spectral methods. *Econometrica*, 37(3):424–438, 1969.

[9] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *J. Finance*, 49(5):1639–1664, 1994.

[10] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.

[11] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*, 2012.

[12] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

[14] I. Laptev and T. Lindberg. Space-time interest points. In *ICCV*, 2003.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[16] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report*, 2009.

[17] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: action recognition through the motion analysis of tracked features. In *ICCV Workshop*, 2009.

[18] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[19] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, pages 1–11, 2012.

[20] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[21] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[22] C. Schuldt, I. Laptev, and C. B. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, 2007.

[24] F. Shi, E. Petriu, and R. Laganiere. Sampling strategies for real-time action recognition. In *CVPR*, 2013.

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2013.

[26] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013.

[27] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[28] S. Wu, B. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *CVPR*, 2010.

[29] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, 2011.

[30] Y. Yamashita, T. Harada, and Y. Kuniyoshi. Causal flow. In *ICME*, 2011.

[31] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

[32] S. Yi and V. Pavlovic. Sparse granger causality graphs for human action classification. In *ICPR*, 2012.

[33] Y. Zhou, S. Yan, and T. S. Huang. Pair-activity classification by bi-trajectories analysis. In *CVPR*, 2008.