# Categorizing Turn-Taking Interactions

Karthir Prabhakar and James M. Rehg

Center for Behavior Imaging and RIM@GT
School of Interactive Computing, Georgia Institute of Technology
{karthir.prabhakar,rehg}@cc.gatech.edu

**Abstract.** We address the problem of categorizing turn-taking interactions between individuals. Social interactions are characterized by turn-taking and arise frequently in real-world videos. Our approach is based on the use of temporal causal analysis to decompose a space-time visual word representation of video into co-occuring independent segments, called *causal sets* [1]. These causal sets then serves the input to a multiple instance learning framework to categorize turn-taking interactions. We introduce a new turn-taking interactions dataset consisting of social games and sports rallies. We demonstrate that our formulation of multiple instance learning (QP-MISVM) is better able to leverage the repetitive structure in turn-taking interactions and demonstrates superior performance relative to a conventional bag of words model.

## 1 Introduction

The categorization of activities from video data is a long-standing problem in computer vision, with numerous applications in areas such as video understanding, multimedia retrieval, and surveillance. Historically, much attention was focused on the activities of an individual subject, often using simplified video content [2,3]. As a consequence of improved feature representations [4], there has been significant progress in categorizing actions using realistic video footage [4,5]. However, the majority of the actions in these recent datasets still involve a single actor and are of relatively short duration. In this paper we address the categorization of turn-taking interactions between two or more individuals over extended time periods. We introduce a novel dataset of turn-taking activities which includes both YouTube videos of social games between children and parents as well as sports rallies taken from broadcast footage.

A basic challenge of activity recognition in unstructured real-world footage is the presence of a wide variety of clutter and distractors. Existing video categorization methods based on space-time visual words (STVW) encode the video volume holistically, incorporating all of the detected STVW into a global representation. In reality, an activity of interest will occupy only a portion of the video volume, and the motion of the camera in conjunction with independent background motions will generate numerous spurious features. This issue is likely to be especially problematic in the case of long video sequences containing complex activities.

One solution to this problem is to segment the video. Standard approaches to video segmentation utilize motion and appearance features to segment video into regions or volumes at the pixel level. In general, however, the automatic segmentation of general

video content remains challenging. For the case of turn-taking videos, however, we have recently proposed a segmentation method [1] which groups space-time visual words (STVW) into non-interacting *causal sets* based on their temporal co-occurrence. This segmentation approach is effective in retrieving turn-taking interactions. However, it is not obvious how to leverage this segmentation for the problem of supervised categorization:

1. The basic segmentation method depends upon the choice of a threshold parameter to identify which words should be grouped together, but it is not clear how to automatically determine the thresholds for good supervised categorization performance.
2. While we have a category label for each video clip, we don't know which causal sets contain the activity. For supervised learning, how can we predict category labels given a segmentation without having labels for the segments?

We demonstrate that both of these challenges can be successfully addressed using a novel formulation of multiple instance learning (MIL).

In this paper, we develop a general method for categorizing video content which contains turn-taking activities, and we make three primary contributions. First, we show that the number of unique causal sets (segmentations) generated by the method of [1] is linear in the number of STVW, which makes it possible to integrate the search for the correct thresholds into the model-fitting process. This automates the previously manually search for good thresholds in [1]. Second, we introduce a variation of multiple instance learning, called QP-MISVM, which can more accurately infer the labels of the causal sets during learning. Third, we present a novel Turn-Taking Activities dataset, consisting of 1,096 video clips of social games and sports rallies take from YouTube and broadcast footage, respectively. This dataset contains a diverse collection of video clips exhibiting substantial clutter and complex interactions. We experimentally validate our method against the holistic STVW approach and standard MIL, and we demonstrate superior performance. We also show state-of-the-art performance on the KTH dataset.

## 2 Related Work

**Recognizing Interactions** There has been little prior work in recognizing interactions between individuals when compared to related areas such as single-person action recognition or human-object interaction recognition. Both [6] and [7] require background subtraction and rely on tracking of body parts, and use very constrained data. In [8], the focus is on short, multi-person interactions such as kick or push in uncluttered environments. The work in [9] uses video sequences from TV shows to recognize two-person interactions such as hand-shake or high-five. They rely on detection of upper-body and estimation of head orientation, which can be difficult to obtain in extended activities. These prior works focus on activities of short duration and don't contain turn-taking. Both [10] and [1] focus on turn-taking interactions, but their task is for retrieval and not categorization.

**Multiple Instance Learning (MIL)** There has been some recent work in applying MIL to the detection and recognition of human actions. In [11], MIL is used to learn sign

language from weakly aligned subtitles. In [12], MIL is used to categorize mouse actions. [13] and [14] use MIL for feature selection and categorize human actions. In [15], MIL is used for action detection in crowded scenes where candidate detections are treated as bags of instances. In [16], MIL is used to overcome noisy labels in YouTube videos for categorization, where the noise comes in incorrectly tagging a video. In this work, both video metadata (title, keywords, etc.) and visual features are used to obtain features for MIL. In contrast to this prior work, we use MIL to simultaneously obtain segmentation labels and categorization in realistic video footage with clutter. To our knowledge, ours is the first work in using MIL for segmentations in clutter.

## 3 Activity Categorization Using Causal Sets

Our goal is to categorize activities which contain turn-taking interactions in natural settings. Since the video sequences of these interactions are obtained from YouTube (social games) and broadcast footage (sports rallies), they contain a wide variety of clutter. We use our temporal causal analysis method [1] to compute a scalar measure of causal influence between each pair of visual words. We demonstrate that the number of possible unique causal sets is linear in the number of visual words, and this makes it possible for us to directly search for the correct graph structure without the need for a pre-defined threshold. In order to leverage the segmentations in categorization, we formulate the problem in a Multiple Instance Learning (MIL) framework and propose a novel variation of MIL, called QP-MISVM, which more accurately infers the labels of causal sets for turn-taking interactions.

### 3.1 Temporal Causal Analysis

We briefly review the procedure of temporal causal analysis from [1] for an example sequence. A video sequence is discretized by extracting space-time interest points [4], and each interest point $p$ has an associated feature vector $f_p$ with two components: position-dependent histograms of oriented gradients (HoG) and optical flow (HoF) from $p$'s space-time neighborhood. We build spatio-temporal visual words by applying k-means clustering to $\{f_p\}$'s and assigning each interest point to its closest visual word. An example sequence of a pattycake game with the visual words overlayed is shown in Fig. 1. For simplicity of discussion, we have only shown a small subset (five) of the visual words from the sequence. In addition to the pattycake interaction, the sequence contains two children in the background who move around independently of the interaction, and the visual words corresponding to them are shown in brown and yellow.

(a) Frame 156 (b) Frame 163 (c) Frame 181

**Fig. 1.** Visual words from a pattycake game. Each color corresponds to a visual word.

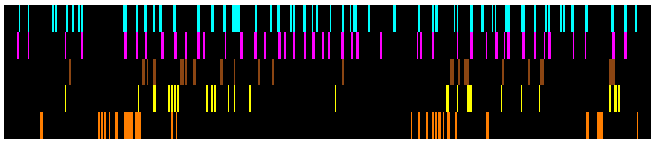
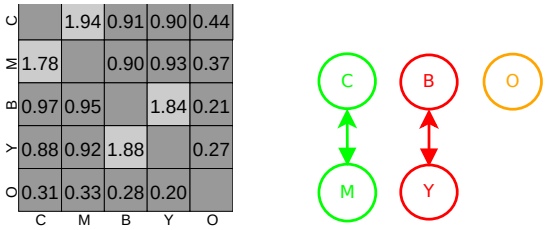


**Fig. 2.** Point-processes for the visual words in Fig. 1.

Each visual word is represented as a temporal point process by marking the times of occurrence of the visual word in a video. The point processes corresponding to the five visual words in the pattycake example are shown in Fig. 2. In this example, we notice that the processes form three sets based on their temporal co-occurrence: the cyan-magenta processes (pattycake game), the brown-yellow processes (children in background), and the orange process (camera motion). A measure of influence between processes is obtained using a pair-wise test for Granger causality, resulting in a causal score $C(i, j)$ which encodes influence of process $i$ on process $j$. Fig. 3(a) shows the causal score matrix for the point processes in Fig. 2. We notice that the pairwise causal scores between cyan and magenta and between brown and yellow are high, while causal scores for the orange process (background) is much lower. To identify the existence of Granger causality from the causal score matrix, a statistical null-hypothesis test based on surrogate data is employed to threshold the causal score matrix, resulting in a segmentation of the visual words into *causal sets*. A set of causal sets generated from a particular threshold is called a *causal collection*. The causal collection corresponding to a threshold at 95% significance level for the pattycake example is shown in Fig. 3(b), and the overlayed visualization of the causal collection is shown in Fig. 4.




(a) Causal Score Matrix (b) Causal Graph

**Fig. 3.** Causal analysis of point processes in Fig. 2. In 3(a), causal scores above the threshold are shown in a lighter color.

(a) Frame 156 (b) Frame 163 (c) Frame 181

**Fig. 4.** Visualizations of causal sets in Fig. 3(b)

### 3.2 Multiple Instance Learning for Causal Sets

The causal sets derived from temporal causal analysis do not naturally lend themselves to standard supervised learning approaches. Some of the causal sets correspond to the interaction and the rest to background, but the label for a causal set is unknown during training as well as testing. The simplest solution is to manually label each of the causal sets as foreground or background, as was done in [1] for categorization of social games. However, this approach is cumbersome and does not scale.

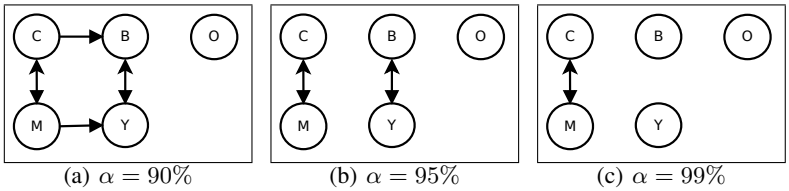
We make two observations which enable us to solve this problem. First, we do not need to obtain labels for all of the causal sets in order to perform categorization. For each video sequence, if we can find one causal set which contains a good segmentation of the interaction, which we call the *representative set*, then the problem is reduced to that of a standard supervised learning problem. A good segmentation contains most of the interaction and very little of the clutter. Second, the causal sets can be thought of as instances in a multiple instance learning (MIL) problem, where the label of the bag (video) is known but the labels of the instances (causal sets) are unknown. The terms causal-sets and instances are used interchangeably. In the MIL framework, the labels for the instances can be automatically inferred from the labels of the bag.

We evaluate two different MIL approaches. First, we use the maximum bag margin formulation of MIL (MI-SVM) [17] where a single instance is chosen as the representative instance for the bag. This is useful since we only need to obtain the representative instance. Second, we use multiple instance learning via embedded instance selection (MILES) [18] which transforms the multiple-instance problem to that of a standard supervised problem without the need to relate particular instances to their labels. This is done by mapping a bag into a feature space defined by the instances in the training set.

There are two problems that must be addressed in applying multiple instance learning to our task. First, the identification of causal sets relies on a choice of threshold, and we do not know how to choose the threshold for best performance. Second, videos in natural settings will contain many more negative instances (clutter) than positive instances (interaction). In MI-SVM, the initial average representation of instances in the bag is not representative of the activity in the video because of high number of negative instances (clutter). In MILES, the mapping of bag to all the instances in the training set does not result in a sparse similarity matrix, which MILES relies on, because of high number of negative instances. In the following sections, we address these concerns. First, we propose a representation for bags from causal sets to address the challenge with picking good thresholds. Second, we extend both MI-SVM and MILES to utilize the turn-taking nature of the activities in selecting relevant instances.

### 3.3 Bag Representation for Causal Sets

The natural representation of a bag is the video sequence, and the instances within a bag are the causal sets derived from temporal causal analysis. For the pattycake example in Fig. 3(b), there are three instances corresponding to the three causal sets: cyan-magenta, blue-yellow, and orange. These particular causal sets were obtained by thresholding the causal score matrix at a 95% significance level. A different choice of a threshold value




(a) $\alpha = 90\%$ (b) $\alpha = 95\%$ (c) $\alpha = 99\%$

**Fig. 5.** Sensitivity to thresholding.

will produce a possibly different causal collection. Fig. 5 illustrates 3 different causal collections for the pattycake example with their associated confidence levels.

Without the knowledge of what the particular causal sets represent in a video sequence, it is unclear which causal collection produces the correct segmentation of the video. In general a single choice of significance level (e.g. 95%) will not be appropriate for all videos. We need a representation of a bag where the choice of instances (causal sets) are independent of a threshold value so that at least one of the instances is guaranteed to contain a good segmentation of the interaction. A naïve choice of a bag representation that would guarantee this is to include all possible combinations of visual words as instances, but this will include many noisy instances in the positive bag (e.g. combination of cyan-orange or magenta-yellow-orange in the pattycake example). Another possibility would be to generate all possible subgraphs of the causal score matrix as the instances of a bag. There are $2^{k^2}$ possible subgraphs for a graph with $k$ vertices, where $k$ is the number of visual words. Even for a small $k$ this is expensive. We make two key observations that constrain the problem:

**Observation 1** *For a $k \times k$ causal score matrix, we have $(k^2 - k)/2$ threshold values that affect the partition. In order to change a causal set, it is necessary to remove the causal links in both directions between two visual words. This requires a threshold greater than $max(C(i,j), C(j,i))$.*

Then, the number of causal collections for a $k \times k$ causal score matrix is $(k^2 - k)/2$ since we can convert the causal score matrix to a symmetric matrix by $C(i,j) = \max(C(i,j), C(j,i))$ without changing the partition. This is much lower than the number of possible subgraphs, but the total number of instances in the set of all causal collections is still high. However, in addition, we observe that:

**Observation 2** *By organizing threshold values such that $thresh_i < thresh_{i+1}$, $\forall i \in \{1, ..., k^2 - k\}$, we notice that $|C_{thresh_{i+1}}| = |C_{thresh_i}| + 1$, where $C_{thresh_i}$ is the causal collection at $thresh_i$. Since only one new instance is added for each threshold value, there will be at most $2k - 1$ instances for a $k \times k$ causal score matrix.*

As a consequence of these observations, it is feasible to exhaustively enumerate the unique causal sets based on sorting the entires in the causal-score matrix. For the pattycake example, the bag representation is:

$$\begin{aligned}\mathbf{B} = \{&(C), (M), (B), (Y), (O), (C, M), (B, Y),\\ &(C, M, B, Y), (C, M, B, Y, O)\}\end{aligned}$$

### 3.4 Feature Representation

For each instance in a bag, we build a spatio-temporal bag-of-features (BoF) representation by constructing a visual vocabulary with k-means from features sampled in the training videos. Each feature in a given instance is assigned to its closest vocabulary-word, and we compute a histogram of visual word occurrences over a space-time volume corresponding to various spatio-temporal grids, following the procedure in [4]. Then each bag for a video sequence is represented as a set of histograms of visual words in each causal set for each channel, where each channel is a combination of a spatial and temporal grid along with either a HoG or HoF descriptor.

### 3.5 Quasi-Periodic MI-SVM

In the MI-SVM formulation, a classifier is initially trained on the average representation of all the instances in each positive bag and all of the negative instances in each negative bag. The margin of the positive bag is defined by the margin of the *most positive* instance, and the margin of a negative bag is defined by the margin of the *least negative* instance. The instances in the positive bags are evaluated against the learned decision function, and the instance which maximizes the decision value in each positive bag is chosen as the new representative instance of the bag. This process is repeated until convergence, when the assignments of the representative instances do not change.

However, such a representation is not robust when there are many negative instances in the bag that can skew the initialization. In this section, we extend MI-SVM to leverage the repetitive structure in turn-taking interactions, via the notion of quasi-periodicity from [10]. In the next section, we will extend MILES similarly. We previously used quasi-periodic scores to retrieve social-games in [1]. The heuristic scoring function described in [10] measures the degree of quasi-periodicity in a video sequence by converting each video frame to a symbol and analysing the resulting string pattern. It can be applied to a single causal set by restricting the analysis to the visual words that the set contains.

For each instance in a bag, we represent each frame of the sequence by the histogram of visual words that are contained in the instance, and assign an event label $e$ to each frame by applying k-means clustering to the histogram representation of frames. For each event $e$, we define event information as $I(e) = -\log_{|E|} p(e)$, where $|E|$ denotes the total number of events and $p(e)$ is frequency of $e$ in the sequence. For each pattern, we compute pattern information $I(\text{pat})$ by computing the sum of unique events $u$ in pat as $I(\text{pat}) = \sum_i I(u_i)$. Then, we compute the quasi-periodic score for the pattern as:

$$G(\text{pat}) = I(\text{pat}) * (\text{Occur}(\text{pat}) - 1) \tag{1}$$

where $\text{Occur}(\text{pat})$ is the number of occurrences of pat. We compare the quasi-periodic scores against a minimum-gain (e.g. min-gain $= 1$) and only accept patterns which exceed this measure. The quasi-periodic pattern score for some of the instances in the pattycake example are: $G(\text{pat}_{(C,M)}) = 4.80$, $G(\text{pat}_{(B,Y)}) = 3.59$, $G(\text{pat}_{(C,M,B,Y)}) = 4.99$, and $G(\text{pat}_{(O)}) = 0.41$.

The quasi-periodic scores by themselves are not enough to discriminate the causal sets containing an interaction from the causal sets containing background. In many

**Algorithm 1**: QP-MISVM

**Input**: $\mathcal{B}$ (set of bags) , $X$ (train data) , $Y$ (train labels)
**Output**: w,b
`/* for instances in positive bags */`
**foreach** c $\in \mathcal{B}_+$ **do**
  $Pat_c \longleftarrow$ GetPat(c)
  evaluate G($Pat_c$) from Eq. 1
$I \longleftarrow$ instances with G($Pat_c$) $>$ threshold
initialize $\mathbf{X}_I \longleftarrow \frac{1}{|I|}\sum_{i\in I}\mathbf{X}_i$
**repeat**
  compute SVM solution $\mathbf{w}, b$ for data $\mathbf{X}_I, Y_I$
  compute outputs $f_i =< \mathbf{w}, \mathbf{X}_{+,i} > + b, \forall i \in I$
  $s(I) \longleftarrow \arg\max_{i\in I} \alpha f_i + \beta \mathrm{G}(Pat_i), \forall i \in I$
  $\mathbf{X}_I \longleftarrow \mathbf{X}_{s(I)}$
**until** *selection variables* $s(I)$ *have not changed* ;

videos, background or camera motions recur throughout the sequence, producing natural repetitiveness (e.g. the $(B, Y)$ causal set in the pattycake example). These spurious motion patterns can be segmented from true interactions through temporal causal analysis. But their quasi-periodic scores are above the min-gain and they cannot be identified as background instances by quasi-periodic analysis alone (e.g. see Figs. 9-14(b)).

We can leverage the quasi-periodic scores in a discriminative framework by biasing the MI-SVM initialization and update steps with those patterns that have high quasi-periodic scores. For each instance in a given bag, we compute it's quasi-periodic score from Eq. 1. We initialize all the positive bags by averaging over all the instances in the bag that have their quasi-periodic score above min-gain. This biases the initialization towards patterns which have repetitive structure. The initialization for the negative bags is same as the MI-SVM formulation. During each iteration, the representative instance for the positive bag is chosen as the instance which gives the maximum value for the linear combination of the learned decision function and the quasi-periodic score:

$$I^* = \arg\max_{c\in\mathbf{B}} \alpha f_c + \beta G(Pat_c) \tag{2}$$

where $c$ is the set of causal sets in the bag $B$, $\alpha$ and $\beta$ are the mixing values. In our experiments, $\alpha$ and $\beta$ are initialized to 0.5. For classification, we use a non-linear support vector machine with a multi-channel $\chi^2$ kernel, similar to [4].

### 3.6 Quasi-Periodic MILES

In the MILES formulation, each bag is embedded into the instance-space by measuring the similarity between each instance $x^k$ in the training set and a bag $B_i$, and it is defined by the closest instance in the bag. A bag is then mapped into the instance-space. The intuition behind this embedding is that if some instance $x^i$ achieves high similarity to


**Fig. 6.** Turn-Taking Interactions Dataset: Sports Rallies

some positive bags and low similarity to some negative bags, then the feature $x^i$ is useful in categorizing the bags. This mapping will result in a sparse embedding since only certain instances will have high similarity to bags of some category. However, an issue with such a representation in our task is that there are many more negative instances than positive instances, and the resulting embedding will not be sparse.

We leverage the quasi-periodic scores to extend MILES for our task. Instead of using all the instances for embedding a bag, we create a small subset of concept class instances $c^m$ which correspond to instances in the training set which have high quasi-periodic scores. Then, we embed a bag in the space of this new concept class similarly as the original definition [1]:

$$s(c^m, B_i) = max_j \mathbf{exp}\left(-\frac{||x_{ij} - c^m||^2}{\sigma^2}\right) \tag{3}$$

and the embedding is now into the smaller concept class space:

$$B'_i = [s(c^1, B_i), \ldots, s(c^m, B_i)]^T \tag{4}$$

## 4 Empirical Evaluation

In this section, we show experimental evaluation of our proposed approaches on 3 different datasets. In addition to our two proposed approaches QP-MILES and QP-MISVM, we also evaluate several intermediate approaches. We evaluate the performance of using QP-Scores alone to choose salient causal groups. That is, we choose all causal groups in a video which has a QP-score above some threshold. We choose the threshold empirically during the training and validation stage. Also, we evaluate the performance of using the standard MI-SVM and MILES algorithms without incorporating the QP heuristics. Finally, we further consider a baseline approach: a single-instance

[1] The pseudocode for QP-MILES is available at http://www.cc.gatech.edu/cpl/projects/temporalcausality.


**Fig. 7.** Turn-Taking Interactions Dataset: Social Games

SVM where all detected interest points are used to generate single feature vector for the video sequence. The baseline was motivated by [4]. Our goal is to empirically evaluate each step of our approach and to understand how each element adds to the approach.

### 4.1 Turn-Taking Interactions Dataset

We introduce a new challenging dataset consisting of two types of turn-taking interactions: social games and sports rallies. These interactions constitute our Turn-Taking Interactions (TTI) dataset, which has 1,096 video segments in total. We collected social interactions in 5 different categories from YouTube: baby toss, ball roll, pattycake, peekaboo, and tickle. Many of the videos are from home movies of a parent-child interaction with some child-child and adult-adult interactions. Since the videos were collected in natural settings, there is considerable intra-class variability in how a particular social game is played, and this can be seen in the sample frames from the dataset of social-games in Figs. 7. There are considerable variations in view-points within the same class, and there is also considerable camera movement and noise from background motion (see Figs. 10-14).

We also collected video sequences of 4 rally-based sports from broadcast footage. A video segment is labeled as a rally if the ball passed at least twice between the players. Sample frames from our dataset of sports-rallies are shown in Fig. 6. The challenge in the sports rallies is not only the intra-class variation in viewpoint and appearance of the rally but also the similarity across the rallies in appearance and viewpoint. For example, many of the tennis and table-tennis rallies have similar appearance. Due to the nature of broadcast footage, there are also considerable camera movement (panning and zooming) and scene cuts.

| Methods | [2] | [4] | [5] | [19] | [20] | QP-Scores | QP-MILES | QP-MISVM |
|---|---|---|---|---|---|---|---|---|
| Accuracy | 71.7% | 91.8% | 93.8% | 94.7% | 96.7% | 93.5% | 94.7% | 96.3% |

**Table 1.** KTH Comparisons

### 4.2 Results on KTH Human Actions

We evaluated the performance of our method on the KTH actions dataset [2]. While none of the KTH sequences contain turn-taking interactions, natural repetition occurs within the sequences (e.g. waving and camera zooming several times), enabling the use of our method. We follow the experimental setup of [2] for training and testing. Tab. 1 compares our average class accuracy to previously published results. Overall, we achieve performance which is consistent with the state-of-the-art. See Tab. 3 for a detailed comparison of our QP-based approaches, and Fig. 9 for a visualization.

| | Baseline | Hand-Labeled | **QP-MISVM** |
|---|---|---|---|
| Babytoss | 45.9% | 57.9% | 52.4% |
| Pattycake | 49.4% | 72.1% | 58.1% |
| Tickle | 45.3% | 56.8% | 47.8% |

**Table 2.** Comparisons between QP-MISVM and hand-labeled segments.

### 4.3 Results on Manually Hand-Labeled Examples

We now compare the performance of our automated approach to our earlier results based on hand-labeling. In [1], we presented preliminary categorization results in which the causal sets corresponding to the interaction were manually identified. These hand-labeled segments provide the best possible training data for an activity, and their performance can be thought of as an upper-bound for any automated approach. In Table 2, we compare the performance of QP-MISVM to both the baseline and the hand-labeled case. We achieve an overall accuracy of 52.8%, while the baseline and hand-labeled results reported in [1] were 46.8% and 62.3%, respectively.

| | Baseline | MI-SVM | MILES | QP-Scores | **QP-MILES** | **QP-MISVM** |
|---|---|---|---|---|---|---|
| KTH | 91.8% | 90.3% | 87.0% | 93.5% | 94.7% | 96.3% |
| Social Games | 58.9% | 62.2% | 53.6% | 60.1% | 63.4% | 71.7% |
| Sports Rallies | 64.1% | 70.8% | 57.4% | 72.8% | 70.0% | 80.7% |

**Table 3.** Comparisons between different proposed approaches.

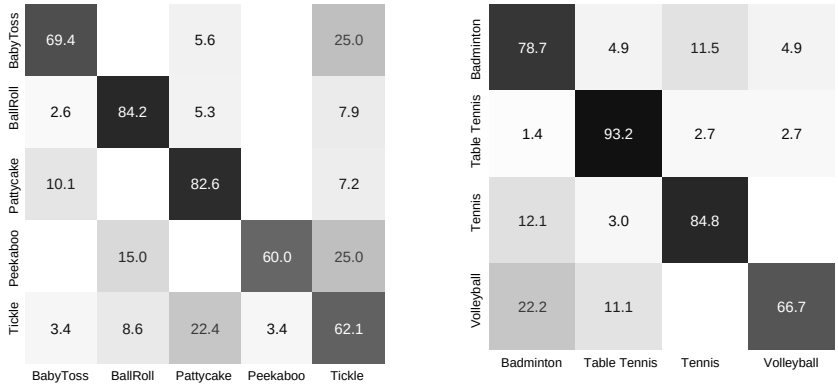
### 4.4 Results on Turn-Taking Interactions

We evaluate our performance in categorizing turn-taking interactions. We split the videos into equal training and testing sets and adopt a one-against-all approach to multi-class classification. For the sports rallies, we organize the training set and testing set such that rallies from the same match do not overlap across the training and testing sets.

Results from our approaches are reported in Table 3. Overall, QP-MISVM achieves the best performance overall across the dataset. It is clear that adding QP-heuristics improved performance in both MILES and MI-SVM. However, the performance of

QP-MILES is close the performance of QP-Scores, and there is not an added benefit by incorporating the MILES framework. However, QP-MISVM clearly outperforms all the approaches by a significant margin. The confusion matrices for the QP-MISVM approach are shown in Fig. 8.

The visualizations from QP-Scores and QP-MISVM are shown in Figs. 10-14. In the babytoss example, the camera motion generates spurious features on the background, and in the pattycake example, there are spurious features from both camera motion and from the two children in the background. For both the tennis and volleyball examples, the spurious features are generated by panning and zooming of the camera, which is natural in broadcast footage. The badminton example is particularly challenging because there is a scene-cut which shows a smash. For all the sequences we see that QP-Scores discards many of the clutter, but it still contains clutter. This is because the clutter occurs throughout the video, and thus contain repetitive structure. However, QP-MISVM extracts visual words that compactly represent the activity and discards most background clutter.

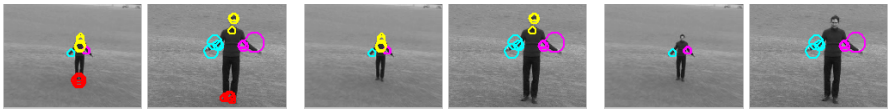



(a) Social Games (Avg. 71.7%) (b) Sports Rallies (Avg. 80.7%)

**Fig. 8.** Confusion matrices on Turn-Taking Interactions dataset.

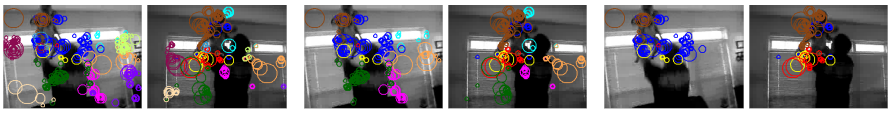
## 5 Conclusion

We present an approach to categorizing videos of turn-taking activities which leverages the segmentation from temporal causal analysis. Our approach is based on a variant of MIL which can automatically label the causal sets as foreground/background while simultaneously predicting the activity category label. We showed that QP-MISVM achieves superior performance over the baseline, MI-SVM, MILES, QP-Scores, and QP-MILES.



(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

**Fig. 9.** Handclapping with scale variation



(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

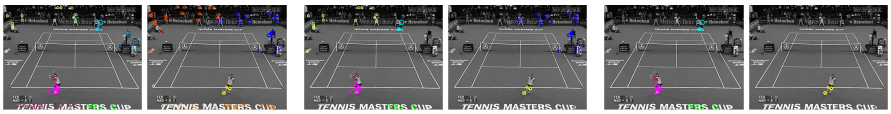
**Fig. 10.** Babytoss sequence with camera motion

(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

**Fig. 11.** Pattycake sequence with camera movement and children in the background

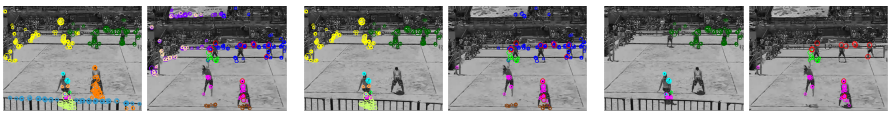
(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

**Fig. 12.** Badminton sequence with scene-cut



(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

**Fig. 13.** Tennis sequence with camera movement



(a) All Visual-Words (b) QP-Scores (c) QP-MISVM

**Fig. 14.** Volleyball sequence with camera motion

## 6 Acknowledgements

Portions of this research were supported in part by NSF Award IIS-1016772. In addition, the first author was supported by an NSF Fellowship.

## References

1. Prabhakar, K., Oh, S., Wang, P., Abowd, G.D., Rehg, J.M.: Temporal causality for the analysis of visual events. In: CVPR. (2010) 1, 2, 3, 5, 7, 11
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR. (2004) 1, 10, 11
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV. (2005) 1
4. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1, 3, 7, 8, 10
5. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR. (2009) 1, 10
6. Datta, A., Shah, M., Lobo, N.D.V.: Person-on-person violence detection in video data. In: ICPR. (2002) 2
7. Park, S., Aggarwal, J.: Simultaneous tracking of multiple body parts of interacting persons. CVIU (2006) 2
8. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activites. In: ICCV. (2009) 2
9. Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.: High five: Recognising human interactions in tv shows. In: BMVC. (2010) 2
10. Wang, P., Abowd, G.D., Rehg, J.M.: Quasi-periodic event analysis for social game retrieval. In: ICCV. (2009) 2, 7
11. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching tv (using weakly aligned subtitles). In: CVPR. (2009) 2
12. Nguyen, M., Torresani, L., Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV. (2009) 3
13. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. PAMI (2010) 3
14. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. In: ECCV. (2010) 3
15. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.: Action detection in complex scenes with spatial and temporal ambiguities. In: ICCV. (2009) 3
16. Leung, T., Song, Y., Zhang, J.: Handling label noise in video classification via multiple instance learning. In: ICCV. (2011) 3
17. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS. (2002) 5
18. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-instance learning via embedded instance selection. In: PAMI. (2006) 5
19. Kaaniche, M., Bremond, F.: Gesture recognition by learning local motion signatures. In: CVPR. (2010) 10
20. Schindler, K., van Gool, L.: Action snippets: how many frames does human action recognition require. In: CVPR. (2008) 10