Spatio-Temporal Human-Object Interactions for Action Recognition in Videos

Victor Escorcia and Juan Carlos Niebles
Electrical and Electronics Engineering Department
Universidad del Norte, Colombia
{escorciav, njuan}@uninorte.edu.co

Abstract

We introduce a new method for representing the dynamics of human-object interactions in videos. Previous algorithms tend to focus on modeling the spatial relationships between objects and actors, but ignore the evolving nature of this relationship through time. Our algorithm captures the dynamic nature of human-object interactions by modeling how these patterns evolve with respect to time. Our experiments show that encoding such temporal evolution is crucial for correctly discriminating human actions that involve similar objects and spatial human-object relationships, but only differ on the temporal aspect of the interaction, e.g. answer phone *and* dial phone *We validate our approach on two human activity datasets and show performance improvements over competing state-of-the-art representations.*

1. Introduction

Current algorithms for human activity recognition in videos can achieve promising classification performance in benchmark tests [10, 13, 19]. However, it is still challenging for most algorithms to generate semantic descriptions or achieve detailed understanding due to the broad gap between the algorithmic representation and the high-level semantics. The standard baseline algorithm for action recognition, which uses a visual bag-of-features representation from spatio-temporal interest points [5, 10], can categorize simple actions (*e.g*. stand-up, hand-shake) with moderate success, but has trouble capturing semantic cues such as the objects involved in an action or the relationships between objects and actors. In this paper we are interested in representations of human actions that can capture semantic cues in order to reduce the lack of description inherent in many traditional approaches.

A promising direction to augment the level of semantic detail in action understanding is that of modeling human-object interactions. Human-object interactions can be a powerful cue that provides contextual information about actors and objects and has been shown to be critical for successful action recognition in still images [22]. Current methods for modeling interactions in video, focus on capturing object co-occurrence and relative spatial locations between objects and actors [6, 18]. These cues may be enough when each action of interest involves a different object, *e.g*. dial phone and drink, or when actions that involve the same object have distinctive relative locations between the object and the person manipulating it, *e.g*. tennis forehand and tennis serve. However, we are usually interested in several actions that involve the same object and may only be distinguished by modeling how the object interacts with

Figure 1. Objects play a valuable role to describe and categorize human actions as they provide strong cues about which actions are taking place. However, recognizing the objects involved in an event may not be a sufficient cue for discrimination. In the example above, it is not enough to recognize a phone to distinguish between *answer phone* (top row) and *dial phone* (bottom row). Furthermore, modeling their overall relative spatial configuration does not provide discrimination power, as the object is always placed in similar configurations with respect to the actor. In order to discriminate these actions, it is crucial to model how the human-object relationship *changes* through time.

the actor across time, *e.g*. dial phone and answer phone. Unfortunately, current algorithms are unable to encode this information.

In this paper, we address this issue by introducing a method for encoding cues related to the spatio-temporal interaction between objects and humans in video. Our algorithm combines information about object and actor, their relative locations and the evolution of the interaction through time. By aggregating these cues, our algorithm is able to capture subtle differences in actions that only differ in the temporal evolution of the human-object interaction, such as dial phone and answer phone. Additionally, we show our framework yields a descriptive representation from which we can learn semantically meaningful cues that improve action recognition performance.

The rest of the paper is organized as follows. Section 1.1 overviews some of the related work. Section 2 describes our proposed spatio-temporal representation of human-object interactions. Section 3 presents our human action recognition framework that leverages the proposed representation. We present experimental validation in Section 4 and conclude the paper in Section 5.

1.1. Related Work

Human activity recognition is one of the most active areas of computer vision. Here, we only review some of the most relevant aspects of the literature and refer the reader to recent surveys [1] for broader coverage.

An important goal in activity recognition systems is to automatically obtain detailed interpretations of the behaviors and events in a scene. While current algorithms are still far from achieving this, some recent work have started to incorporate more detailed semantic cues. One set of methods attempt this by introducing intermediate representations can be related to semantic interpretations. These intermediate representations may be viewed as semantic spaces on which we can project videos and obtain meaningful descriptions. Choices of spaces with promising performances are attributes [12], attributes and parts [23] or simple action detectors in an action bank framework [20]. Unfortunately, we usually require highly supervised training data in order to construct these semantically meaningful representations.

Other methods adopt human-centric representations to encode cues that are more relevant to the activities of interest. Usually, these algorithms perform tracking [17, 8] as a preprocessing step in order to localize the actor in the scene and represent the visual and motion patterns using local video features [8]. Alternatively, one could treat the localization of the actor as a latent variable in a discriminative framework [9], with the price of increased representation and model complexity.

Also along this direction, and closely related to our work, some researchers have proposed the introduction of human-object interaction cues, which refer to the relationship between actors and objects involved in an action. These cues were originally introduced to address the problem of action recognition in still images, and include information about static cues such as relative location of the object and human, relative sizes, poses, among others [23]. In the video domain, some authors have [3, 6, 18] introduced algorithms to capture human-object interactions for action recognition. However, these representations still focus on spatial relationships only and lack the ability to capture the temporal evolution of the human-object interaction through time.

In this paper, we are interested in incorporating cues related to the spatio-temporal dynamics of human-object interactions in video. Therefore, our work is also related to methods that encode temporal structures or temporal evolution of cues for activity recognition. These include methods that model spatio-temporal relationships of visual features such as hierarchical spatio-temporal segmentations encoded in graph structures [2], low-level features grouped into temporal [15, 16] or spatio-temporal grids [4, 10], as well as traditional sequential models such as HMMs [7], CRFs [21] or DBNs [11].

2. Representing Spatio-Temporal Human-Object Interactions

In this section we introduce our representation for human-object interactions in videos. We propose a representation that captures cues related to the relationships between humans and objects, which we illustrate in Figure 2. Given the locations of a human $H_i^{1:T}$ and an object $O_j^{1:T}$ in a video sequence of length T, the goal of our representation is to encode information about how the interaction between human and object evolves through time. As shown by our experiments, describing the evolution of the interaction helps discriminating actions that: (i) involve similar objects, so that object occurrence is not a sufficient cue for discrimination; (ii) involve objects and humans that maintain similar spatial relationships, so that a global measure of the relative locations and sizes is not a powerful cue; (iii) may only be discriminated by analyzing the temporal aspect of the relationship between object and actor.

To encode the temporal evolution of the interaction, we aggregate information about the relative location, sizes of objects and humans into a spatio-temporal grid. Intuitively, the choice of temporal quantization instead of explicitly modeling the interaction with a continuous dynamical model provides robustness to noise and intra-class variance.

In practice, our representation aggregates the following cues from every frame t in the sequence:

Relative location ϕ_l^t: encodes the relative spatial location of the object O_j^t with respect to the human H_i^t in frame t. We encode this relationship using an elliptical binning in

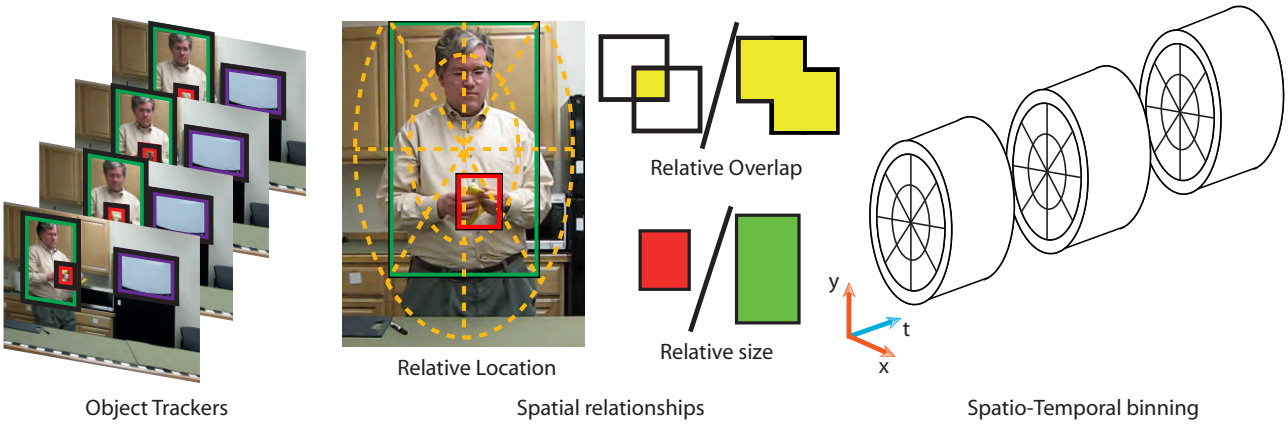

Figure 2. Our descriptor encodes the dynamics of spatio-temporal human interactions in videos. The descriptor works on pairs of human and object tracks (first column) and computes features (second and third column) related to: relative object location with respect to the human, overlap between human and object, and their relative sizes. In order to capture how the interaction between human and object evolves through time, our method aggregates these features into time intervals. As an example, we illustrate the case of 3 non-overlapping intervals (fourth column), which define a spatio-temporal binning where relative location information is aggregated.

the image frame as illustrated in Figure 2. Note that we adopt a soft-voting strategy, where the object casts votes with weights that are inversely proportional to the distance to the center of the bin. The soft-voting strategy allows to deal with uncertainty in the object shape and location. The elliptical binning takes into account the aspect ratio of the human body and softens the bounding boxing location of the person. As illustrated in Figure 2, in practice we use 8 angular and 2 radial partitions as well as 1 bin for objects outside the largest radius. This yields a 17 dimensional descriptor.

Relative sizes ϕ_r^t: encodes the ratio of the area in pixels between the human and object windows, *i.e.* $\phi_r^t = |O_j^t|/|H_i^t|$. This feature is helpful to implicitly define constraints about the size of the object, *e.g.* the cup is smaller than the human in drinking actions.

Relative overlap ϕ_o^t: this feature is computed as the intersection area over union area between human and object windows. This feature combines cues about the size and distance of the object with respect to the human in a single scalar.

We compute ϕ_l^t, ϕ_r^t and ϕ_o^t at each frame t in the time interval $t = [1, T]$. In order to encode how these features evolve through time, we aggregate them at several time intervals. In general, we define a set $\mathcal{V}$ of V time intervals $I_v = [t_v^{start}, t_v^{end}]$. Each interval v is associated with a feature vector $\Phi^v = [\Phi_l^v, \Phi_o^v, \Phi_r^v]$, by aggregating features as follows:

$$\Phi_l^v = \sum_{t \in I_v} \phi_l^t \tag{1}$$

$$\Phi_o^v = \left[\begin{array}{ccc} \max_{t \in I_v} \phi_o^t & \min_{t \in I_v} \phi_o^t & \overline{\phi_o^t} \end{array} \right] \tag{2}$$

$$\Phi_r^v = \left[\begin{array}{ccc} \max_{t \in I_v} \phi_r^t & \min_{t \in I_v} \phi_r^t & \overline{\phi_r^t} \end{array} \right] \tag{3}$$

Finally, we simply concatenate the descriptors extracted at each interval into a single vector that describes the interaction over the entire sequence:

$$\Phi(H_i, O_j) = \left[\Phi^1, \Phi^2, \ldots, \Phi^V\right] \tag{4}$$

Our representation can be used with several choices of the set $\mathcal{V}$. Natural choices are: (a) splitting a video into non-overlapping temporal intervals of equal length (Figure 2), (b) temporal pyramid with video intervals of multiple time scales. We study both options experimentally in our empirical evaluation in Section 4.

Figure 3 visualizes our computed features for two example videos of *making a phone call* and *answering a phone call*. We note the use of multiple time intervals (top row) yields descriptors with higher discrimination power in comparison to global aggregation (bottom row).

3. Action Classification

In this section, we describe how we can use our representation for human-object interactions to perform human activity recognition in videos. We describe the steps for training classifiers for each action of interest from data, as well as for recognizing actions in new sequences.

3.1. Training

The goal of the training stage is to learn a classifier for each action of interest. Our system takes input videos that contain examples of the action category. These videos are annotated with: the temporal subsequence where the action

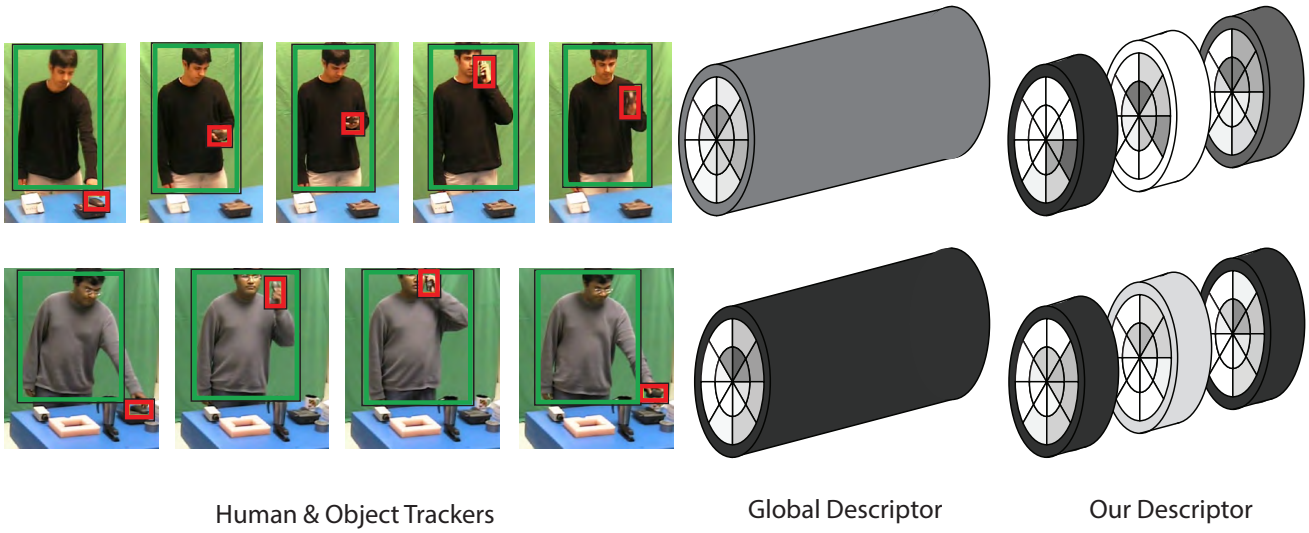

Figure 3. Spatio-temporal human-object interactions. Given tracks of objects and humans in input videos (first column), our descriptor computes features related to the relationship between object and actor. Competing methods aggregate these features into a single spatial grid (second column), ignoring all the temporal information related to the interaction. Our descriptor aggregates this information into multiple time intervals separately. In this example, our descriptor (third column) aggregates information into 3 non-overlapping intervals of equal length that cover the entire sequence. This enables our algorithm to leverage the temporal evolution of the human object interaction, providing a performance boost in human action recognition tasks.

occurs, and bounding boxes for the object and actor performing the action in at least one frame.

As a first pre-processing stage, our system needs to track the person and object during the temporal subsequence that contains the action. This may be done using a combination of low-level trackers and object detectors as in [18]. The goal of this step is to provide the system with localizations of the object and actor at each frame, which are required to capture the dynamics of the interaction between human and object. In practice, we run a correlation-based low-level tracker from the annotated bounding boxes. This simple tracker may fail in some videos, so we manually add a few annotations until the low-level tracker produces an acceptable track. Note that we choose to provide relatively good tracks to our descriptor in order to isolate the effects of tracker failure and focus our evaluation on the discriminative power of our proposed descriptor.

Given the human and object tracks produced by our simple tracker, we compute descriptors for all human-object pairs in the training set. We use the features Φ introduced in Section 2, which capture information about the relative location of human-object, their sizes and temporal evolution of these relationships.

Finally, we ensemble a training set by computing features from positive and negative examples for each class. In practice, our positive examples are descriptors computed from human-object pairs associated with an action label, while any human-object pair associated with other labels are taken as negatives. We train a discriminative classifier to learn the decision boundaries between action categories in the interaction feature space. In practice, we choose a linear SVM framework to learn these boundaries.

3.2. Recognition

At recognition time, our algorithm localizes and recognizes human actions in novel sequences.

Similar to training, we run a tracking pre-processing stage to obtain multiple human and object tracks. This process tends to produce many tracks, and the task of our classifier consists of discriminating which human-object pairs truly correspond to each action of interest.

The algorithm forms candidate human-object pairs by grouping tracks that are close in space and time. For each candidate pair, we compute the corresponding features Φ as introduced in Section 2.

Finally, we compute confidence scores with our discriminative human action classifier and declare a human-object pair as positive if it is above certain decision threshold. In a multi-class classification setting, we test each human-object pair against all human action classifiers, we make a classification decision for each such pair by choosing the class with highest confidence score.

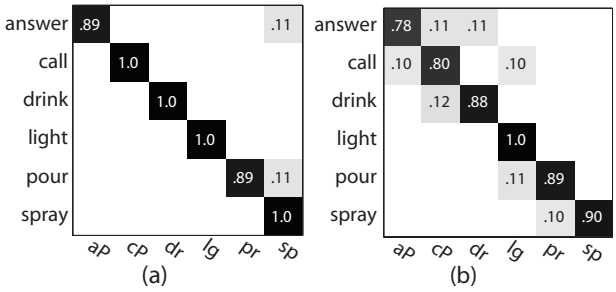

Figure 4. Confusion matrices on the Gupta action dataset. (a) Our spatio-temporal human-object interaction descriptor, (b) human-object interaction descriptor from [18].

Method	Accuracy	meanAP
Global HOI descriptor [18]	87.5 %	91.9 %
(a) Global Relative Location (Φ_l)	82.5 %	80.2 %
(b) Global Interaction (Φ_l, Φ_s, Φ_o)	90.1 %	83.5 %
(c) Spatio-Temporal Relative Location	88.5 %	87.3 %
(d) Spatio-Temporal Interaction	**96.3 %**	**93.2 %**

Table 1. Quantitative evaluation of different components of our descriptor on the Gupta action dataset. (a) aggregates the Φ_l over a single time interval that covers the entire sequence. (b) combines (Φ_l, Φ_s, Φ_o) over a single time interval that covers the entire sequence. (c) and (d) aggregate features over 3 non-overlapping temporal intervals of equal length that cover the entire sequence.

4. Experimental results

To validate our approach, we test our human action recognition algorithm in action classification tasks using two public datasets [6, 14].

Gupta action dataset First, we evaluate our method using the multi-class human action dataset proposed by Gupta *et al*. [6]. This dataset contains 10 actors performing 6 actions that involve 4 object classes for a total of 54 videos. The actions in this dataset are: drinking from a cup, spraying from a bottle, answering a phone call, making a phone call, pouring from a cup and lighting a flashlight. The videos are taken in laboratory conditions, using a static camera in a scene with green background and white objects.

We use a 4-fold cross-validation experimental setting to evaluate our algorithm in this dataset. We train our human action classifier using spatio-temporal interaction features as described above. In all cases, we train binary classifiers by using examples from other classes as negatives. During the recognition step, all human-object pairs found in a novel video are scored under each of the 6 action models, so that the predicted action corresponds to the classifier with highest confidence score over all pairs.

In order to focus the evaluation on the power of our proposed descriptor, we use the same input human and object tracks when comparing multiple methods. We summarize the results of our quantitative evaluation in the confusion matrix of Figure 4(a), whose diagonal averages to 96.3%. We also compare our system to our implementation of the state-of-the-art human-object interaction descriptor of Prest *et al*. [18], which computes interactions and relationships over a single temporal interval and does not capture the temporal evolution of the interaction. The associated confusion matrix is presented in Figure 4(b), whose diagonal averages to 87.5%. We note that our descriptor provides better discrimination, especially since the actions of interest involve similar objects and spatial relationships that can only be discriminated using the temporal aspect of the interaction.

Choices for the set $\mathcal{V}$	Accuracy	meanAP
Spatio-temporal pyramid, 3-levels	85.2 %	**95.5** %
Flat, 1 global interval	92.3 %	83.5 %
Flat, 3 non-overlapping intervals	**98.1 %**	93.2 %

Table 2. Experimental comparisons of different choices of $\mathcal{V}$

Method	Accuracy	meanAP
Global HOI descriptor [18]	90.7 %	93.9 %
Global Relative Location	82.0 %	92.7 %
Global Interaction	92.7 %	89.5 %
Spatio-Temporal Relative Location	92.7 %	90.1 %
Spatio-Temporal Interaction	**98.0 %**	**94.2 %**

Table 3. Quantitative evaluation of different components of our descriptor on the Rochester Daily Activity dataset. See Table 1 for more details.

Furthermore, we evaluate the contribution of each component of our descriptor to the final recognition performance. Table 1 summarizes our quantitative comparisons for each component under two measures: Accuracy, which corresponds to the average of the confusion matrix diagonal in multi-class setting; and meanAP, which corresponds to the mean Average Precision for each binary classifier. We note that each feature component is complementary, and we can achieve the best performance by combining our relative location, sizes, overlap and spatio-temporal binning.

Table 1 also compares our descriptor against a baseline that accumulates interaction cues in a temporally global manner. This is similar to the descriptor proposed in [18]. Here, we report figures obtained with our own implementation of [18], which uses the input human-object tracks obtained as described above. This enables a fair comparison against our proposed descriptor.

We also study the effect of using different choices of the set $\mathcal{V}$, which we summarize in Table 2. We note that using a spatio-temporal pyramid affects performance in terms of accuracy, which we attribute to overfitting due to the higher dimensionality of the descriptor and the small training set.

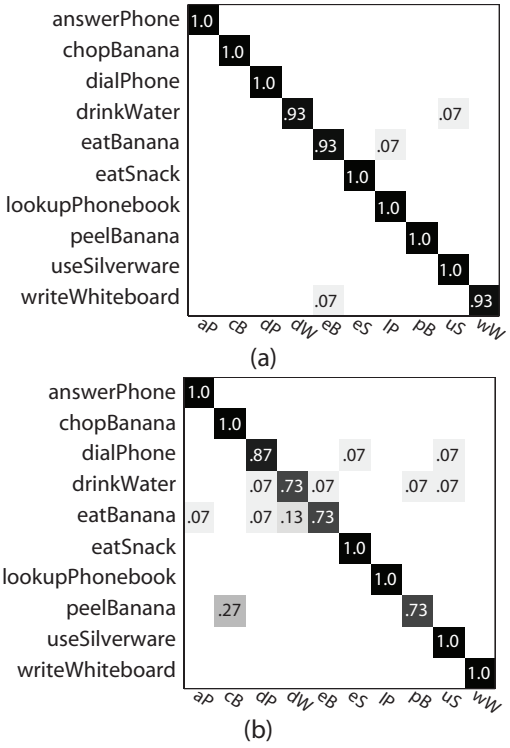

Figure 5. Confusion matrices for the Rochester Daily Activity dataset. (a) Our spatio-temporal human-object interaction descriptor, (b) human-object interaction descriptor from [18].

Rochester Daily Action dataset In a second experiment, we evaluate our algorithm using the Rochester Daily Action dataset [14]. This dataset contains 5 actors performing 10 actions that involve 8 object classes for a total of 150 videos. In comparison to the Gupta dataset, this is a more challenging dataset with videos recorded in a real kitchen scenario and real objects.

We summarize the recognition performance of our algorithm with the confusion matrix in Figure 5(a). We also compare it to our implementation of the state-of-the-art descriptor from [18], whose confusion matrix is in Figure 5(b). Similarly to our previous experiment, we note that our descriptor has a higher discrimination power in comparison to the best interaction descriptor reported in the literature [18]. We also provide comparisons to our baselines in Table 3.

Finally, Figure 6 shows examples of successful action recognition as well as classification errors made by our system. We note most errors are due to strong similarities of spatio-temporal relationships between human and objects. We intend to address this issue in the future with a more structured description of human-object interactions that allow the occurrence of multiple objects.

5. Conclusions and Future Work

This paper presents a new descriptor for human object interactions that captures the dynamics and temporal evolution of the relationship between object and human. Our experimental evaluation shows that this descriptor can provide better discrimination power for the task of human action classification in comparison to state-of-the-art algorithms. We plan to study the inclusion other contextual cues, such as location, pose and motion of multiple surrounding objects and actors.

Acknowledgements

This work was funded by a Google Research Award. V.E. is supported by Colciencias Young Scientist and Innovator Fellowship. J.C.N. is supported by a Microsoft Research Faculty Fellowship.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis. *ACM Computing Surveys*, 43(3):1–43, Apr. 2011.

[2] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, pages 778–785. IEEE, Nov. 2011.

[3] C.-Y. Chen and K. Grauman. Efficient activity detection with max-subgraph search. In *CVPR*, pages 1274–1281. IEEE, 2012.

[4] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Proc. of 9th International Workshop on Visual Surveillance (VSWS09) in conjuction with ICCV*, 2009.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *VSPETS*, pages 65–72. IEEE, 2005.

[6] A. Gupta and L. S. Davis. Objects in Action: An Approach for Combining Action Understanding and Object Perception. In *CVPR*. IEEE, June 2007.

[7] N. Ikizler and D. A. Forsyth. Searching for Complex Human Activities with No Visual Examples. *IJCV*, 80(3):337–357, 2008.

[8] A. Klser, M. Marszaek, C. Schmid, and A. Zisserman. Human Focused Action Localization in Video. In *International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV*, 2010.

[9] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, pages 2003–2010. IEEE, Nov. 2011.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, volume 1. IEEE, 2008.

[11] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *CVPR*, pages 1–8. IEEE, June 2007.

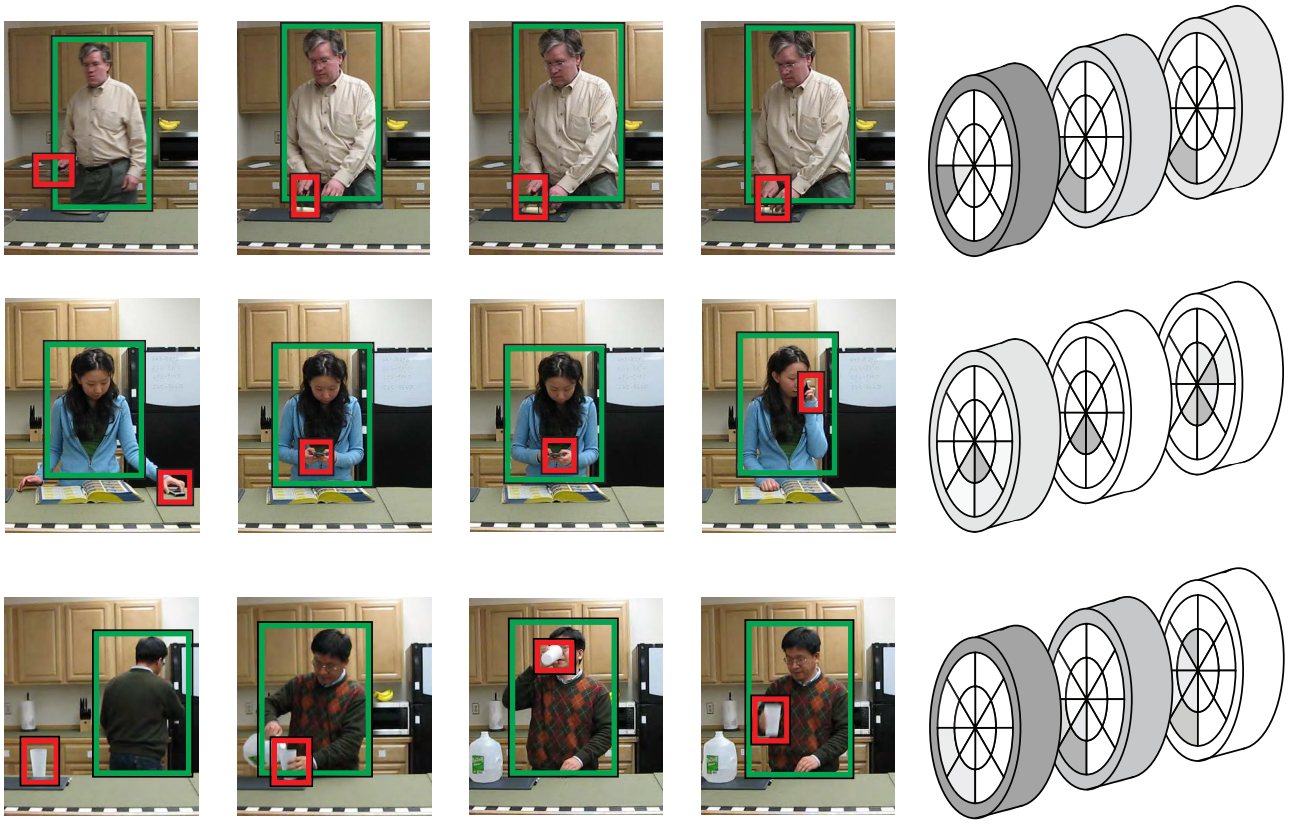

Figure 6. Example recognition results. The top 2 rows show examples of successful action recognition in the Rochester Daily Action dataset. We note that in both examples, our spatio-temporal human-object interaction descriptor correctly captures the evolution of the interaction across time. The bottom row shows an example of a mistake made by our classifier. In this case, the spatio-temporal relationship between the actor and the cup is similar to the typical interaction between the actor and a fork. We intend to solve this issue in the future using a richer description of the human-object interactions.

[12] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344. IEEE, June 2011.

[13] J. Liu, J. Luo, and M. Shah. Recognizing Realistic Actions from Videos in the Wild". In *CVPR*, volume 2, pages 1996–2003. IEEE, 2009.

[14] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, Washington, DC, USA, 2009. IEEE Computer Society.

[15] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *ECCV*, volume 6312. Springer, 2010.

[16] B. Packer, K. Saenko, and D. Koller. A combined pose, object, and feature model for action understanding. In *CVPR*, pages 1378–1385. IEEE, June 2012.

[17] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12):2441–2453, 2012.

[18] A. Prest, V. Ferrari, and C. Schmid. Explicit Modeling of Human-object Interactions in Realistic Videos. *IEEE TPAMI*, pages 9–16, Aug. 2012.

[19] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *CVPR*. IEEE, 2008.

[20] S. Sadanand and J. Corso. Action Bank: A High-Level Representation of Activity in Video. In *CVPR*, 2012.

[21] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *CVIU*, 104(2-3):210–220, Nov. 2006.

[22] B. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In *CVPR*. IEEE, 2010.

[23] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*. IEEE, Nov. 2011.