"Important Stuff, Everywhere!" Activity Recognition with Salient Proto-Objects as Context

Lukas Rybok, Boris Schauerte, Ziad Al-Halah, and Rainer Stiefelhagen

Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
{name.surname}@kit.edu

Abstract

Object information is an important cue to discriminate between activities that draw part of their meaning from context. Most of current work either ignores this information or relies on specific object detectors. However, such object detectors require a significant amount of training data and complicate the transfer of the action recognition framework to novel domains with different objects and object-action relationships. Motivated by recent advances in saliency detection, we propose to employ salient proto-objects for unsupervised discovery of object- and object-part candidates and use them as a contextual cue for activity recognition. Our experimental evaluation on three publicly available data sets shows that the integration of proto-objects and simple motion features substantially improves recognition performance, outperforming the state-of-the-art.

1. Introduction

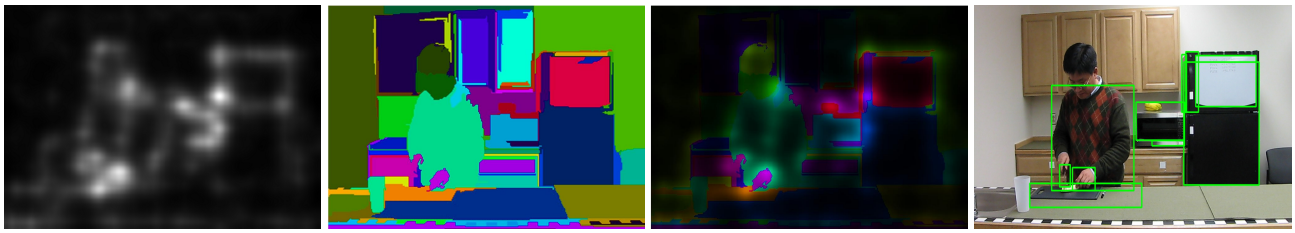
To recognize actions and activities is an important component of many computer vision applications such as, for example, human-robot interaction, surveillance, and multimedia retrieval. While many approaches are designed to classify simple actions – *i.e.*, "motion events" – such as "standing up" or "walking" [21], the focus of our work lies on the recognition of complex action sequences that are also known as "activities".

Following action identification theory [36], an action (and thus activity) derives its meaning from the context and not from the motion alone. Such contextual information may involve, among others, the state of the actor's mind, the location where the action takes place, as well as the objects that are manipulated by the actor. However, most works in action and activity recognition ignore contextual cues and focus on the identification of activities based on motion patterns alone (*c.f.* [1, 39]). On the other hand, approaches that do incorporate object information usually depend on detectors that require supervised training (*e.g.* [35, 16]). Since the detectors require a substantial amount of manually annotated training data, expanding such frameworks (*e.g.*, adding new action classes) becomes the bottleneck for generalized tasks. As an alternative, we propose to use proto-object features, which do not require any supervision, as a contextual cue for activity recognition.

Attention forms a selective gating mechanism that decides what will be processed at later stages (*e.g.*, object recognition) and is often described as a "spotlight" that enhances the processing in the attended, *i.e.*, "illuminated", region [26]. Interestingly, experimental evidence suggests that attention can be tied to objects, object parts, and/or groups of objects [7, 31]. But, how can we attend to objects before we recognize them [37]? Rensink introduced the concept of proto-objects in his coherence theory [30, 37] and defined them as volatile units of visual information that may be validated as actual objects through focused attention. In other words, proto-objects are object- or object-part candidates that have been detected, but not yet identified.

Motivated by the ability of humans to reliably determine such visually salient regions from the background, many approaches have been proposed to detect proto-objects with the least statistical knowledge of the objects themselves, *e.g.*, [14, 37, 4, 12, 33]. Since visual attention and object recognition are tightly linked processes in the human visual system, there is an increasing interest in integrating both concepts to increase the performance of computer vision systems. For instance, Walther and Koch [37] combine an attention based system with SIFT-based object recognition and demonstrate that such an integration can improve the overall performance. Other applications involve the prediction of human gaze patterns [33], scene understanding [12], and object detection [2].

In this work, we show that proto-object detection allows us to find object candidate regions that can be used as a cue for motion based activity recognition. We evaluate the proposed features in combination with a simple bag-of-words model [18] on three challenging data sets and demonstrate



(a) Saliency map (b) Image segmentation (c) Saliency-weighted segments (d) proto-object locations

Figure 1. Overview of the proto-object detection approach. First, a QDCT-based saliency map and a graph-based image segmentation are calculated. Then, the segments with the highest saliency are selected as object candidates. This illustration is best seen in color.

that our approach greatly increases the recognition rate. This way, we are able to improve the state-of-the-art on the KIT Robot-Kitchen [32], URADL [20], and CAD-120 [16] data sets by 0.5%, 2.0%, and 4.3%, respectively.

2. Related work

In recent years, action and activity recognition in videos have received an increasing amount of attention within the computer vision community [21, 1, 39]. However, the majority of works neglects contextual information and focuses on the recognition based on motion patterns alone [20, 18, 38, 42, 19].

Some simple ways to incorporate object knowledge is by directly using ground-truth labels [17], and possibly adding artificial noise to simulate imperfect detections [10] or by attaching RFID Tags to all relevant objects [40]. Approaches that solely use image features to automatically retrieve object information mostly rely on trained detectors [16, 11, 35, 25, 27, 6]. But, building a robust detector handling all types of object classes is still challenging and subject for future research.Thus, for each new domain a dedicated detector needs to be trained instead, which requires time and cost expensive data collection and annotation. Furthermore, since different object states (*e.g.*, *opened* vs. *closed* fridge) also contain meaningful information, the detectors also need to discriminate between states with high accuracy.

To circumvent such shortcomings, approaches that can automatically extract potentially relevant image regions have been proposed. Ikizler-Cinbis and Sclaroff [13] assume large moving regions as candidate objects. Similarly, Packer *et al*. [23] rely on background subtraction in combination with articulated body tracking and assume that candidate objets are foreground regions that cannot be explained by limbs. Furthermore, the hand regions are included in the set of object candidates, since the hand can easily occlude large parts of smaller objects. This way, most of the objects that are being manipulated by the observed person can be obtained. However, unlike in the presented approach, objects that remain static throughout most of the activity are left unnoticed (*e.g.*, the majority of the dishes located on the table during the activity "eating dinner").

One automatic approach to mine discriminative image regions based on their strong correlations with the target class is the grouplet [41]. However, this feature representation draws its power from grouping codewords and preserving their spatial configurations. Nonetheless, in activities of daily living (*e.g.*, kitchen related activities) the spatial relations between different objects involved in one activity are often arbitrary.

Most related to our approach is the work of Prest *et al*. [28], which employs an objectness measure to determine the most relevant region that is located close to the actor. However, unlike our approach, their method is relatively complex, only operates on still images, depends on human detection [2], and only considers the object that is part of the interaction.

3. Proto-Object Extraction

In the following, we describe how we use proto-objects as object candidates to enrich motion descriptors with contextual image information for activity recognition. We build our framework upon Schauerte and Stiefelhagen's (see [33]) quaternion-based spectral saliency detection algorithm. Among the advantages of this approach are its simplicity, theoretical soundness, high accuracy in predicting foreground regions, and that it is fully unsupervised. The algorithm extends Hou *et al*.'s [12] "image signature" descriptor by employing a quaternion representation of an image, which makes it possible to process all color channels simultaneously in a holistic fashion.

Hou *et al*.'s image signatures are defined as the signum function of the Discrete Cosine Transform (DCT) of an image. A saliency map can be obtained by applying an inverse DCT to an image signature followed by smoothing with a Gaussian kernel [12, 33]. It has been demonstrated theoretically and experimentally that this approach concentrates the image energies on foreground regions [12]. We calculate the saliency maps based on the CIE L*A*B color space, since it has been shown to reliably yield better performance than most other color spaces [33].

Peaks in a saliency map only indicate the positions of

Figure 2. Representatives of the first 18 proto-object feature codebook entries for subject 1 of the URADL data set. The codewords were selected based upon their Minimal-Redundancy-Maximal-Relevance score [24].

Input: Maximum saliency threshold θ and maximal number of segments to select K
Output: Set of detected proto-objects O
Find maximal saliency value $s_{\max}$;
Set $s' = s_{\max}$; $O = \{\}$;
while $s' > \theta \cdot s_{\max}$ *AND* $|O| < K$ **do**
  Set s' to maximal saliency value;
  Add image segment containing s' to O;
  Set saliency of the selected segment to 0;
end

Algorithm 1: Detecting proto-objects.

the proto-objects, however the approximate spatial extent of each proto-object region still needs to be determined. One common approach is to operate on the saliency map itself, *e.g.*, by region growing or by thresholding [12]. Yet, such a procedure is often highly sensitive to the choice of the saliency detection parameters which directly influences the size of the segmented proto-objects. Instead, we use the saliency map to guide the proto-object selection directly in the image, as shown in Fig. 1. First, we use Felzenszwalb's graph-based algorithm [9] to segment each frame of a video sequence using parameters yielding preferably large image segments (see Fig. 1(b)). In order to select a set of proto-objects, we then apply Algorithm 1, which implements attentional shifts and inhibition of return. In our experiments, we empirically determined its parameters and set $\theta = 70\%$ and $K = 30$.

To encode the appearance of the proto-object regions, we use Dalal and Trigg's Histograms of Oriented Gradients [5], which proved – in preliminary experiments – to be superior to other popular feature descriptors such as, *e.g.*, SIFT, SURF, and ORB. Finally, we apply k-means clustering to obtain a codebook for our proto-object based features. As can be observed in Fig. 2, many of the codewords correspond to real-world objects or object-parts that are meaningful for activity recognition.

4. Activity Recognition

Since object knowledge alone is not enough information to discriminate activities, we employ the well known Space Time Interest Points (STIP) [18] as motion descriptors. To this end, we use Laptev *et al.*'s Harris3d Interest Point detection and as features we either use Histograms of Optical Flow (HOF) alone or in combination with Histograms of Oriented Gradients (HOG). It is noteworthy that the HOG descriptor in this context differs from Dalal and Trigs's original HOG descriptor [5], because it is built through accumulation of gradients within the spatio-temporal cuboid region of a STIP. Thus, it can be seen more as a motion descriptor, since it captures a moving region's change of location, as well as appearance.

A whole image sequence is represented as a bag-of-words, using a 1000-element codebook for motion features (HOF/HOGHOF), and a 200-element codebook for object candidate features (proto-object based features/image segments, see Sec. 3). For simplicity, we use feature fusion via concatenation when combining features from different sources.

To classify a video, we utilize a linear multi-class Support Vector Machine (SVM) [15]. Since it is desirable for subsequent algorithms (*e.g.*, decision fusion or ranking) to provide normalized confidence scores as classification result, we train a multinomial logit model on the training data via cross-validation. This way, we can map the SVM's output (*i.e.*, the distance to the hyperplane) to $[0, 1]$.

It has been pointed out that the power transform of elements in a feature vector F makes the distribution of the features uniform and this way increases the discriminative power of F [3, 29]. Thus, we first apply an L1-normalization to F and then raise each element of F to the power of α. As suggested by Ren and Ramanan [29], we set $\alpha = 0.3$. Finally, all features are standardized to zero-mean and unit-variance, since this feature scaling method proved to yield robust results.

5. Experimental Evaluation

We evaluate our approach on three publicly available benchmark data sets for activity recognition: URADL [20], CAD-120 [16], and KIT Robo-Kitchen [32]. As evaluation measure, we report the overall recognition accuracy. Note that, the lack of training data in CAD-120 prohibit us to robustly learn probability outputs for the classifier, we only report the non-probabilistic version thereof. We focus our evaluation on the aspect of how well the proto-objects perform alone, and combined with motion features (HOF and HOGHOF). To demonstrate the importance of saliency driven object candidate selection, we also compare to the case where all image segments from the segmentation step are used, and, for the URADL data set, the case

of using ground-truth object labels and supervised detectors for selecting object candidates. These segments and candidate objects are described with HOG in the same way as with the proposed proto-object based features. Furthermore, we compare our feature representation with state-of-the art activity recognition approaches to demonstrate its effectivness.

5.1. URADL data set

The University of Rochester Activities of Daily Living (URADL) data set [20] contains 150 high-resolution videos of ten activities which are often similar in motion and thus difficult to be separated without context knowledge. Each activity is performed three times by five different subjects and the evaluation is done using leave-one-person-out cross-validation.

To compare our method with approaches relying on object detections, we annotated all images with the location of the following twelve objects (we will make the annotations publicly available): "whiteboard", "bottle", "cup", "plate", "crisps", "phone", "knifeblock", "paperroll", "phonebook", "peeled banana", "banana". The labels were used to train state-of-the-art object detectors [8] (Mean Average Precision of 0.744), and to determine how well our approach performs compared to using perfect object knowledge. In order to incorporate such object information into our classification framework, we simply use the object classes as codebook entries and calculate bag-of-words histograms.

The results from the experiments and a comparison with state-of-the-art approaches are presented in Tab. 1. Using our proto-objects combined with HOGHOF features yields a perfect recognition accuracy, which is as good as using ground-truth object labels and outperforms state-of-the-art [42] by 2.0%, reaching 100.0%. Also, the use of all image segments without saliency-based region selection and features based on the supervised object detectors perform worse than the proposed method. Furthermore, combining HOF with proto-objects clearly preforms better than combined HOG and HOF features. This suggests that proto-objects are a much better way to capture contextual information than HOG encoded STIP. Surprisingly, HOF with proto-objects also performs better than HOF with ground-truth object labels, which might be because the approach captures more regions that are relevant to the recognition task.

5.2. CAD-120 data set

The Cornel Activity Dataset-120 (CAD-120) [16] contains 120 RGBD videos (we only used the RGB channels) of four subjects performing 10 activities (three repetitions, each time using different objects). Some of the challenges of this benchmark are big variations of camera-view angles and recording locations within each activity class. For comparison, we use the same train-test split that is used in the literature [16] and follow a leave-one-person-out cross-validation protocol.

Method	Accuracy (%)
object detections	68.7
object labels	86.7
all segments	40.0
proto-objects	62.0
HOF	79.3
HOF & object detections	87.3
HOF & object labels	90.0
HOF & all segments	86.7
HOF & proto-objects	97.7
HOGHOF	94.0
HOGHOF & object detections	96.0
HOGHOF & object labels	100.0
HOGHOF & all segments	94.7
HOGHOF & proto-objects	**100.0**
Matikainen *et al.* [19], 2010	70.0
Messing *et al.* [20], 2009	89.0
Prest *et al.* [27], 2012	92.0
Wang *et al.* [38], 2011	96.0
Yi and Lin [42], 2013	98.0

Table 1. Performance results of different methods using a leave-one-person-out testing paradigm on the URADL data set.

The results from the experiments and a comparison with state-of-the-art approaches are presented in Tab. 2. As in the experiments using the other two data sets, it can be observed that combining proto-objects with motion features clearly performs better than using motion features alone. Furthermore, HOGHOF with proto-objects outperforms all other approaches by at least 4.3% (relative improvement), including Koppula *et al.*'s recently proposed state-of-the-art method [16]. The only exception is the work of Koppula and Saxena [17], which however relies on ground-truth object tracks and is thus not comparable to our approach.

The confusion matrix in Fig. 3 reveals that most of the problems of our approach lie in confusing activities including similar motions and objects, such as "microwave", "clean-object" and "take-food", all of which contain interaction with a microwave. To handle this problem, a fine grained motion representation is required, which is left for future work. Still, using proto-objects often helps in such ambiguous situations. For instance, the recognition accuracy of the activities "take-food" and "microwave" increases by 17 and 25 percentage points, respectively, when using HOGHOF with proto-objects compared to HOGHOF.

5.3. KIT Robo-Kitchen data set

The KIT Robo-Kitchen data set (KIT) [32] consists of videos of 14 different activities, each performed by 17 different persons of which ten are used as training data and the remaining seven serve as unseen data for testing. Unlike other benchmarks, one of the challenges of this data set is that the recognition is not based on clips spanning the whole activity, but rather of all possible 150 frame long subsequences of each video. The reasoning behind this is, that

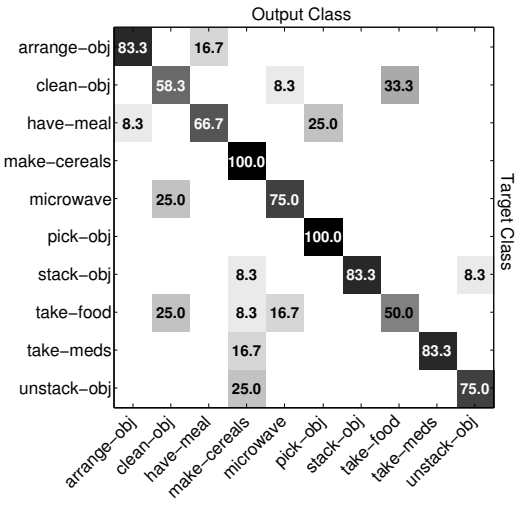


Figure 3. Confusion matrix for the CAD-120 data set when using a combination of HOGHOF and proto-object features.

Method	Accuracy (%)
all segments	40.3
proto-objects	39.5
HOF	66.9
HOF & all segments	71.0
HOF & proto-objects	74.2
HOGHOF	70.2
HOGHOF & all segments	75.0
HOGHOF & proto-objects	**78.2**
Sung *et al*. [34], 2012	26.4
Koppula *et al*. [16], 2013	75.0
Koppula and Saxena [17], 2013	83.1*

Table 2. Performance results of different methods on the CAD-120 data set. *Note that [17] is using ground truth object labels and thus is not directly comparable to our approach.

the data set was designed to model the application of activity recognition in a household robot scenario, in which the robot should offer his services long before the user is finished with the current activity. For a better comparison with other works, we restrict our evaluation on the most popular subset of the data, the setup *room:door*, which consists of ten activity classes.

The results of the experiments are presented in Tab. 3. Here, the combination of HOF with proto-object performs better than all other methods, including the state-of-the-art that is set by Onofri *et al*.'s recent approach [22], which it surpasses by a small margin of 0.5% (relative improvement). It is however surprising, that using proto-object based features alone yields a comparatively high recognition rate. This may be explained with many activities involving objects that are distinctive in their appearance. A clear exception from this are "cut", "peel", which are indeed a major error source. A look at the confusion matrix in Fig. 4 further supports this claim, which backs up the usefulness of proto-object based features as an additional cue for activity recognition.

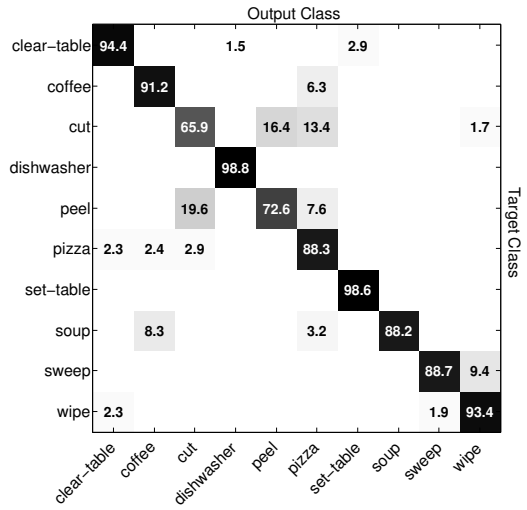


Figure 4. Confusion matrix for the KIT data set when using a combination of HOF and proto-object features.

Method	Accuracy (%)
proto-objects	70.4
HOF	85.6
HOF & proto-objects	**88.7**
HOGHOF	86.6
HOGHOF & proto-objects	88.5
Rybok *et al*. [32], 2011	84.9
Onofri *et al*. [22], 2013	88.3

Table 3. Performance results of different methods on the *room:door* setup of the KIT data set.

6. Conclusion

We propose to use proto-object based features to encode contextual information for activity recognition. The major advantage of our approach is that it allows us to automatically extract object candidates from images without any need for annotated training data or motion information. In an experimental evaluation on three realistic data sets, we showed how well proto-objects complement simple motion features and demonstrated the superior performance over other state-of-the-art approaches. In our future work, we plan to investigate how well a fine-grade motion representation can further help to discriminate between activities involving similar objects and movements.

Acknowledgement. This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human Activity Analysis. *ACM Computing Surveys*, 43(3), 2010. 1, 2

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an Object? In *CVPR*, 2010. 1, 2

[3] R. Arandjelovic and A. Zisserman. Three Things Everyone

Should Know to Improve Object Retrieval. In *CVPR*, 2012. 3

[4] N. D. B. Bruce and J. K. Tsotsos. Saliency, Attention, and Visual Search: An Information Theoretic Approach. *Journal of Vision*, 9(3):1–24, 2009. 1

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 3

[6] V. Delaitre, J. Sivic, and I. Laptev. Learning Person-Object Interactions for Action Recognition in Still Images. In *NIPS*, 2011. 2

[7] J. Duncan. Selective Attention and the Organization of Visual Information. *Journal of Experimental Psychology: General*, 113(4):501–517, 1984. 1

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Rrained Part-based Models. *TPAMI*, 32(9):1627–1645, 2010. 4

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 59(2):167–181, 2004. 3

[10] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen. Combined Intention, Activity, and Motion Recognition for a Humanoid Household Robot. In *IROS*, 2011. 2

[11] A. Gupta, A. Kembhavi, and L. S. Davis. Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. *TPAMI*, 31(10):1775–1789, 2009. 2

[12] X. Hou, J. Harel, and C. Koch. Image Signature: Highlighting Sparse Salient Regions. *TPAMI*, 34(1):194–201, 2011. 1, 2, 3

[13] N. Ikizler-Cinbis and S. Sclaroff. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In *ECCV*, 2010. 2

[14] L. Itti and C. Koch. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. 1

[15] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009. 3

[16] H. S. Koppula, R. Gupta, and A. Saxena. Learning Human Activities and Object Affordances from RGB-D Videos. *IJRR*, 32(8):951–970, 2013. 1, 2, 3, 4, 5

[17] H. S. Koppula and A. Saxena. Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation. In *ICML*, 2013. 2, 4, 5

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *CVPR*, 2008. 1, 2, 3

[19] P. Matikainen, M. Hebert, and R. Sukthankar. Representing Pairwise Spatial and Temporal Relations for Action Recognition. In *ECCV*, 2010. 2, 4

[20] R. Messing, C. Pal, and H. Kautz. Activity Recognition Using the Velocity Histories of Tracked Keypoints. In *ICCV*, 2009. 2, 3, 4

[21] T. B. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *CVIU*, 104(2-3):90–126, 2006. 1, 2

[22] L. Onofri, P. Soda, and G. Iannello. Multiple Subsequence Combination in Human Action Recognition. *IET Computer Vision*, 2013. 5

[23] B. Packer, K. Saenko, and D. Koller. A Combined Pose, Object, and Feature Model for Action Understanding. In *CVPR*, 2012. 2

[24] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *TPAMI*, 27(8):1226–1238, 2005. 3

[25] H. Pirsiavash and D. Ramanan. Detecting Activities of Daily Living in First-person Camera Views. In *CVPR*, 2012. 2

[26] M. I. Posner, C. R. R. Snyder, and B. J. Davidson. Attention and the Detection of Signals. *Journal of Experimental Psychology: General*, 109(2):160–174, 1980. 1

[27] A. Prest, V. Ferrari, and C. Schmid. Explicit Modeling of Human-Object Interactions in Realistic Videos. *TPAMI*, 35(4):835–848, 2012. 2, 4

[28] A. Prest, C. Schmid, and V. Ferrari. Weakly Supervised Learning of Interactions between Humans and Objects. *TPAMI*, 34(3):601–614, 2012. 2

[29] X. Ren and D. Ramanan. Histograms of Sparse Codes for Object Detection. In *CVPR*, 2013. 3

[30] R. A. Rensink. The Dynamic Representation of Scenes. *Visual Cognition*, 7(1-3):17–42, 2000. 1

[31] P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395:376–381, 1998. 1

[32] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen. The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems. In *Humanoids*, 2011. 2, 3, 4, 5

[33] B. Schauerte and R. Stiefelhagen. Quaternion-based Spectral Saliency Detection for Eye Fixation Prediction. In *ECCV*, 2012. 1, 2

[34] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured Human Activity Detection from RGBD Images. In *ICRA*, 2012. 5

[35] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving Bag-of-Features Action Recognition with Non-Local Cues. In *BMVC*, 2010. 1, 2

[36] R. R. Vallacher and D. M. Wegner. What Do People Think They're Doing? Action Identification and Human Behavior. *Psychological Review*, 94(1):3–15, 1987. 1

[37] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, 2006. 1

[38] J. Wang, Z. Chen, and Y. Wu. Action Recognition with Multiscale Spatio-Temporal Contexts. In *CVPR*, 2011. 2, 4

[39] D. Weinland, R. Ronfard, and E. Boyer. A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition. *CVIU*, 115(2):224–241, 2011. 1, 2

[40] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A Scalable Approach to Activity Recognition based on Object Use. In *ICCV*, 2007. 2

[41] B. Yao and L. Fei-Fei. Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions. In *CVPR*, 2010. 2

[42] Y. Yi and Y. Lin. Human Action Recognition with Salient Trajectories. *Signal Processing*, 93(11):2932–2941, 2013. 2, 4