TUHOI: Trento Universal Human Object Interaction Dataset

Dieu-Thu Le
DISI, University of Trento
Povo, 38123, Italy
dle@disi.unitn.it

Jasper Uijlings
University of Trento, Italy
University of Edinburgh, Scotland
jrr.uijlings@ed.ac.uk

Raffaella Bernardi
DISI, University of Trento
Povo, 38123, Italy
bernardi@disi.unitn.it

Abstract

This paper describes the Trento Universal Human Object Interaction dataset, TUHOI, which is dedicated to human object interactions in images.[1] Recognizing human actions is an important yet challenging task. Most available datasets in this field are limited in numbers of actions and objects. A large dataset with various actions and human object interactions is needed for training and evaluating complicated and robust human action recognition systems, especially systems that combine knowledge learned from language and vision. We introduce an image collection with more than two thousand actions which have been annotated through crowdsourcing. We review publicly available datasets, describe the annotation process of our image collection and some statistics of this dataset. Finally, experimental results on the dataset including human action recognition based on objects and an analysis of the relation between human-object positions in images and prepositions in language are presented.

1 Introduction

Visual action recognition is generally studied on datasets with a limited number of predefined actions represented in many training images or videos (Ikizler et al., 2008; Delaitre et al., 2011; Yao and Li, 2010; Yao et al., 2011). Common methods using holistic image or video representation such as Bag-of-Words have achieved successful results in retrieval settings (Ayache and Quenot, 2008). Though these predefined lists of actions are good for many computer vision problems, this cannot work when one wants to recognize *all* possible actions. Firstly, the same action can be phrased in several ways. Secondly, the number of actions that such systems would have to recognize in real life data is huge: the number of possible interactions with all possible objects is bounded by the cartesian product of numbers of verbs and objects. Therefore, the task of collecting images or videos of each individual action becomes infeasible with this growing number. By necessity this means that for some actions only few examples will be available. In this paper we want to enable studies in the direction of recognizing all possible actions, for which we provide a new, suitable human-object interaction dataset.

A human action can be defined as a human, object, and the relation between them. Therefore, an action is naturally recognized through its individual components. Recent advances in computer vision have led to reasonable accuracy for object and human recognition, which makes recognizing the components feasible. Additionally, language can help determining how components are combined. Furthermore, the relative position between human and object can be used to disambiguate different human actions. Perhaps prepositions in natural language can be linked to this relative position between the object and human (e.g., step *out of* a car). To transfer this knowledge from language to vision, it is important that the distribution of the visual actions are sampled similarly as the language data. This requirement is fulfilled when the action frequencies in the dataset mirror the frequencies in which they occur in real life.

To sum up, we aim at building an image dataset which can (1) capture the distribution of human interactions with objects in reality (if an action is more common that the other actions, that action is also observed more frequently in the dataset than the others), (2) provide different ways of describing

[1] Our dataset is available to download at http://disi.unitn.it/ dle/dataset/TUHOI.html

Proceedings of the 25th International Conference on Computational Linguistics, pages 17–24,
Dublin, Ireland, August 23-29 2014.

an action for each image (there are many actions that can be phrased in several ways, for example: fix a bike or repair a bike), (3) help with identifying different verb meanings (for example, the word ‘riding’ has different implications for ‘riding a horse’, ‘riding a car’, and ‘riding a skateboard’).

2 Available image datasets for human action recognition

A common approach to human action recognition is to exploit visual features using bag-of-features or part-based representation and treat action recognition as a general classification problem (Delaitre et al., 2010; Yao and Li, 2010; Wang et al., ; Laptev, 2005). For common actions, it has been shown that learning the joint appearance of the human-object interaction can be beneficial (Sadeghi and Farhadi, 2011). Other studies recognize actions by their components such as objects, human poses, scenes (Gupta et al., 2009; Yao et al., 2011): (Yao et al., 2011) jointly models attributes and parts, where attributes are verbs and parts are objects and local body parts. These studies rely on suitable training data for a set of predefined actions: (Gupta et al., 2009) tests on a 6 sport action dataset, (Yao and Li, 2010) attempts to distinguish images where a human plays a musical instrument from images where he/she does not, (Delaitre et al., 2010) classifies images to one of the seven every day actions, and (Yao et al., 2011) introduces a dataset containing 40 human actions. Most of these datasets were obtained using web search results such as Google, Bing, Flickr, etc. The number of images varies from 300 to more than 9K images. A comparison of the publicly available datasets with respect to the number of actions and their related objects is given in Table 1.

Dataset	#images	#objects	#actions	Examples of actions
Ikirler (Ikizler et al., 2008)	467	0	6	running, walking, throwing, crouching and kicking
Willow (Delaitre et al., 2011)	968	5	7	interaction with computer, photographing, riding bike
Sport dataset (Gupta et al., 2009)	300	4	6	tennis-forehand, tennis-serve, cricket bowling
Stanford 40 (Yao et al., 2011)	9532	31	40	ride horse, row boat, ride bike, cut vegetables
PPMI (Yao and Li, 2010)	4800	7	7	play violin, play guitar, play flute, play french horn
PASCAL (Everingham et al., 2012)	1221	6	10	jumping, playing instrument, riding horse
89 action dataset (Le et al., 2013)	2038	19	89	drive bus, sail boat, ride bike, fix bike, watch TV
TUHOI dataset	10805	189	2974	sit on chair, use computer, ride horse, play with dog

Table 1: A comparison of available human action datasets in terms of number of objects and actions

As can be seen in Table 1, the Stanford 40 action dataset contains quite a big number of images with 40 different actions. This dataset is good for visually training action recognizers since there are enough images collected for each actions divided into training and test sets. There are some dataset in which human action does not involved any object, these actions are for instance running, walking, or actions where objects are not specified such as catching, throwing. These types of actions are not the target domain of our dataset. We aim at recognizing the human object interactions based on objects. With the same object, some actions are also more common than other actions: for example, sitting on a chair is more commonly observed than standing on a chair. We want to capture such information in our dataset which can reflect the human action distributions on common objects, aiming to sample human actions related to objects in the visual world. Furthermore, how actions can be phrased in different ways, or how verbs can have different meanings when interacting with different objects should also be considered. Some actions can only be performed on some particular objects and are not applicable to some other objects: a person can ride a horse, ride a bike, can feed a horse, but cannot feed a bike. This problem of ambiguity and different word uses have been widely studied in computational linguistics, but have received little attention from the computer vision community.

With the aim of creating a dataset that covers these requirements, we collect our dataset starting from images where humans and objects co-occur together and define the actions we observe in each image instead of collecting images for some predefined human actions. This way of annotating actions in images is more natural and helps creating a more realistic dataset with various human actions that can occur in images generally.

Recently, some good works attempted to generate descriptive sentences from images (Farhadi et al., 2010; Kulkarni et al., 2011). In our dataset we focus on human actions, which, if present, are often the main topic of interest within an image. As such, our dataset can be used as an important stepping stone for generating full image descriptions as it allows for more rigorous evaluation than free-form text.

3 TUHOI, the new human action dataset

ImageNet is a hierarchical image database built upon the WordNet structure. The DET dataset in the ImageNet large scale object recognition challenge 2013[2] contains 200 objects for training and evaluation. With the idea of starting from images with humans and common objects, we chose to use this DET dataset as a starting point to build our human action data.

3.1 The DET dataset: Object categories and labels

The 200 objects in the DET dataset are general, basic-level categories (e.g., monitor, waffle iron, sofa, spatula, starfish). Each object corresponds to a synset (set of synonymous nouns) in WordNet. The DET training set consists of single topic images where only the target object is annotated. As such, most images only contain primarily the object of interest and few actions. It is good for learning object classifiers but is not suitable for learning action recognition. In contrast, the validation dataset contains various images where all object instances are annotated with a bounding box. Many of these images contain actions. Therefore we start the annotation from the validation set.

Dataset	#images	#images	#object	#instances/object	#"person"
		having "person"	instances	(min-max-median)	instances
Training	395,909	9,877	345,854	438 - 73,799 - 660	18,258
Validation	20,121	5,791	55,502	31 - 12,823 - 111	12,823

Table 2: The statistics of the DET dataset

As can be seen in Table 2, there are 15,668 images having human and 31,081 human instances in these images. We select only images having human since we want to annotate this dataset with human object interactions. Objects related to clothes such as bathing cap, miniskirt, tie, etc. are not interesting for human actions (most of the time, the action associated with these objects is "to wear"). Therefore, we excluded all these objects from the list of 200 objects above, which are: bathing cap, bow tie, bow, brassiere, hat with a wide brim, helmet, maillot, miniskirt, neck brace, sunglasses, tie.

3.2 Human action annotation

Goal Our goal is to annotate these selected images containing humans and objects with their interactions. Each human action is required to be associated with at least one of the given 200 object categories. We used the Crowdflower, a crowdsourcing service for annotating these images. The Crowdflower annotators are required to be English native speakers and they can use any vocabulary to describe the actions as they wish. Every action is composed of a verb and an object (possibly with a preposition).

Annotation guideline For each image, given all object instances appearing in that image (together with their bounding boxes), the annotator has been be asked to assign all human actions associated to each of the object instance in the image (where "no action" is also possible). Every human actions need to have as object one of the object instances given in that image. For example, if the image has a bike and a dog, the annotator will assign every human actions associated to "bike" and "dog". Every image has been annotated by at least 3 annotators, so that each action in the image can be described differently by different people. Some examples of annotated images in our dataset are given in Figure 1.

3.3 Results of the annotation and some statistics

In total, there are 10,805 images, which have been annotated with 58,808 actions, of which 6,826 times it has been annotated with "no action" (11.6%). On average, there are 4.8 actions annotated for each image (excluding "no action"), of which there are 1.97 unique action/image. Some other statistics of the dataset are given in Table 3: The number of unique verbs per object ranges from 1 (starfish, otter) to 158 (dog). As dogs occur very often in this image dataset (4,671 times), the number of actions associated to it is also larger than other objects.

[2] `http://www.image-net.org/challenges/LSVRC/2013/`

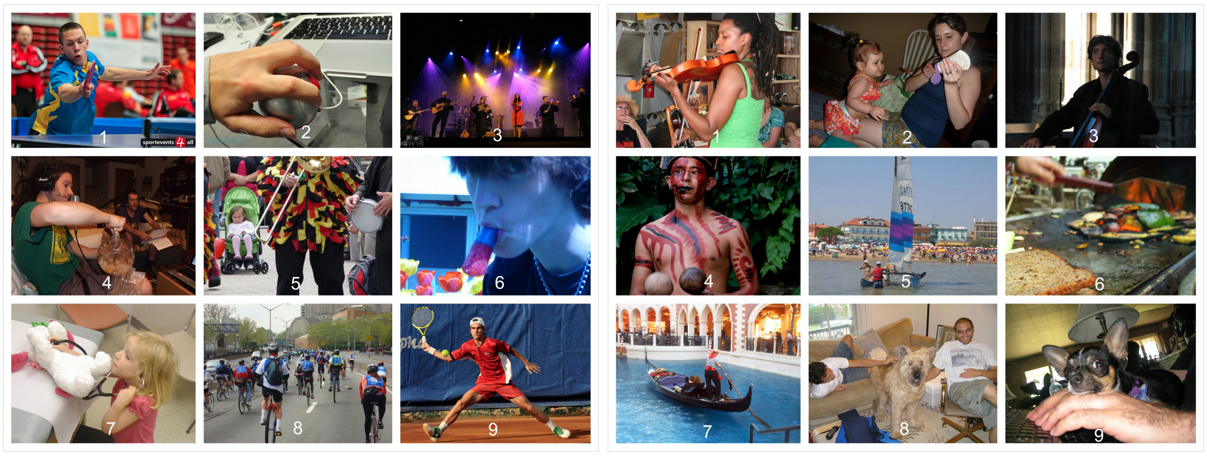


Figure 1: Examples of annotated images: **Left:** (1) play ping-pong, hold racket; (2) use laptop, hold computer mouse; (3) use microphone, play accordion, play guitar, play violin; (4) talk on microphone, sit on sofa, pour pitcher; (5) play trombone; (6) eat/suck popsicle; (7) listen/use/hear stethoscope; (8) ride bicycle, wear backpack; (9) swing/hold racket, hit tennis ball; **Right:** (1) sit on chair, play violin; (2) wear diaper, sit on chair, squeeze/apply cream; (3) sit on chair, play cello; (4) hold/shake maraca; (5) ride watercraft, wear swimming trunks; (6) cook/use stove, stir mushroom, hold spatula; (7) drive/row watercraft; (8) sit on chair, pet dog, lay on sofa; (9) click/type on computer keyboard

Number of unique actions (verb + object): 2,974 actions Number of unique verbs: 860 verbs
Verbs that are used most frequently (verb (#occurrences)): play (13043), hold (7731), ride (4765), sit (3535), sit on (1501) drive (1491), wear (1441), eat (1175), hit (1168), pet (970), use (897), walk (787), stand (756) touch (509), carry (507), blow (384), sail (323), kick (297), lead (290), throw (246), strum (239) stand on (223), run (223)
Verbs that are used least frequently (occur only once): dirty, swing over, twist, beats, walks, ay, curl face, shit, sail in, n‘, see by, forge, draw, tag10, sling, rides, walk across, no image available, waving drag, award, preform, strumb, died, land, unload, tricks, cooked, time, fasten, fall over, holed, leap over, pull up
Objects go with the largest number of verbs (object (#unique verbs)): dog (158), car (80), table (79), watercraft (68) horizontal bar (56), chair (54), cart (52), whale (50), bicycle (48), cattle (42), soccer ball (41), balance beam (38) band aid (38), motorcycle (37), flower pot (35), ladle (35), guitar (35), horse (35), ski (34), bus (34)
Objects that go with the least number of verbs (object (#unique verbs)): milk can (5), pitcher (5), scorpion (4), bear (4), pretzel (4), sheep (4), frog (4), mushroom (4), printer (4), pineapple (4), ruler (3), guacamole (3), isopod (3), chime (3), plate (rack (3), strawberry (3), porcupine (3), ant (3), toaster (3), bagel (3), jellyfish (3), dragonfly (2), lion (2), zebra (2), goldfish (2), hamster (2), fig (2), squirrel (2), bee (2), centipede (2), koala (bear (2), snail (2), pomegranate (2), armadillo (2), otter (1), starfish (1)

Table 3: Some statistics of the human action dataset

For some images, the annotators find many different ways to describe the action in the image. In our data, a set of images was selected to be annotated by more than three people in order to facilitate sanity checks. An example of such image which has been annotated by many people is given in Figure 2. The annotators have found many verbs to describe the action: feeding, leading, running with, touching, giving a treat to, etc.

Splitting training and test set For each object in our human action dataset, we split half of the images for training and the other half is used for testing. The splitting process is done such that actions that occur in test set also occur in training set to guarantee that the training set contains at least one image for each action occurring in the test set.

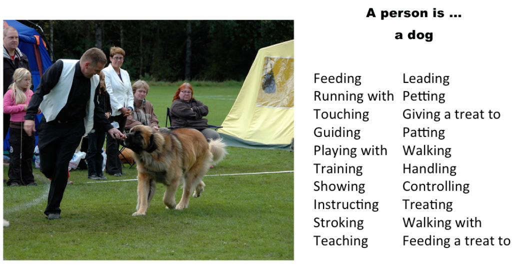


Figure 2: Many different ways to describe an action in an image

Evaluating human action classification in our dataset To evaluate the performance of the human action classification on this dataset, we use two different measurements: the accuracy and the traditional precision, recall and F1 score. The accuracy reflects the percentage of predictions that are correct. We calculate within how many images, the classifier assigns the correct actions for a given object i:

$$Accuracy_i = \frac{\text{number of images that the classifier predicts correctly}}{\text{total number of images}} \tag{1}$$

If the output of the classifier is one of the three annotated actions by human, then the action predicted is considered to be correct. The accuracy of the whole system is the average accuracy over all objects, with n is the total number of objects.

$$Accuracy = \frac{\sum_{i=1}^{n} Accuracy_i}{n} \tag{2}$$

This metric gives us the general performance of the system and easy to interpret. However, it gives higher weights to actions that occur more often in the dataset. For example, if there are many actions "ride bike" occurring in the dataset, the accuracy of the whole system depends mostly on the performance of the class "ride bike". For actions that occur more rarely such as "fix bike", then the accuracy of the class "fix bike" will have little effect to the accuracy of the whole system.

To better analyze the results of the system and evaluate each action individually, we use the precision, recall and F1 score for each class in the classifier. More specifically, as this classifier is the multi-class classifier, these metrics are computed using a confusion matrix:

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}}; Recall_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{3}$$

where M_{ij} is the value of the row i, column j in the confusion matrix. The confusion matrix is oriented such that a given row of the matrix corresponds to the value of the "truth", i.e., correct actions assigned by human, and a given column corresponds to the value of action assigned by the classifier. Finally, the precision, recall and F1 score of the whole system are calculated as the average score over all actions.

4 Experiments

In this part, we use our newly collected dataset for building a general human action classifier based on objects. We analyze the relative positions between humans and objects in each image and use this information to help classifying human actions. Finally, we discuss the relations between human-object positions with prepositions that are used in language for describing human actions.

4.1 Classifying human actions based on human-object positions

In this experiment, we used Forest Random classification method to classify an image to an action given an object. The features used for this classifier are positions of the object and the person appearing in that image. We compare this classifier when using position with a classifier using no position information to see whether position information helps in classifying human actions and in which cases.

Extracting features To extract the features of objects and persons' positions in the images, we take the bounding box of the first object instance annotated in that image. There are images with more than one object instance (for example, there are several 'bike' in an image, so we do not know what 'bike' we are talking about). We use the four coordinates of the bounding boxes of the object and person in the image as features for the classifier.

Results of the classifier To compare whether position information can help in recognizing actions or not, we design a naive classifier which learns from the probability of a verb given an object to assign an action for each image from the training image dataset.

	Accuracy	Precision	Recall	F1
Without position	74.2%	0.40	0.26	0.29
With position	72.1%	0.65	0.29	0.36

Table 4: Results of the classifier with and without position information

Object	Without position	With position	Object	Without position	With position
baseball	0.36	0.52	bus	0.57	0.73
face powder	0.33	1	hair spray	0.73	0.74
harmonica	0.07	0.97	horizontal bar	0.42	0.45
hotdog	0.29	0.57	motorcycle	0.80	0.82
turtle	0.43	0.71	water bottle	0.56	0.65

Table 5: Objects with higher accuracy when using position information

The results of the systems with and without position are report in Table 4. It shows that the accuracy of the classifier without position is higher than when including the position (74.2% in compared to 72.1%). However the precision, recall and F1 of the classifier using position are all higher than without position. It's due to the fact that the classifier without position blindly assigns each image to the most probable action (i.e., actions that occurs most often with a given object learned from the training set), so it obtains better overall accuracy when testing on all images. However, for other possible actions, this classifier is unable to disambiguate actions and the performance of this classifier on less frequent actions is worst than when including position information into the classifier. Generally, when taking into account all possible actions, the position-based classifier has better average precision, recall and F1 score (28.6% without position in compared with 35.8% using position).

To further analyze which objects and actions, the position information helps better, we compare the accuracy of each individual objects. Table 5 reports main objects that have higher accuracy when using position. We want to be able to predict which kind of actions that positions will help in recognizing them through the knowledge we learn from language. This prediction will help us to learn how to include the position information inside our human action recognizer since not all actions can be disambiguated by positions. We divide the actions into two groups: one group for which we found position information increase the classification results. Another group for which we found position information to decrease the classification results.

4.2 From prepositions in language to relative positions between human and object in images

In this section, we want to learn how prepositions in language can be used to determine which positions are useful in action classification, i.e., if they belong to the first group or the second group in the previous experiment.

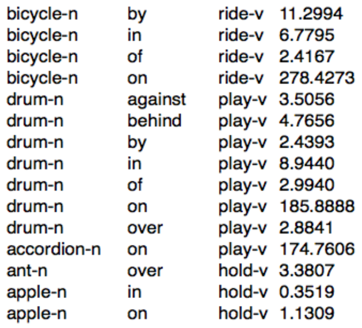
The relative positions between human and object in images are useful in analyzing their interactions. For example, when a person is riding a horse, the person is usually on the top of the horse, and when a person is feeding a horse, then the person is usually standing next to the horse. In spoken English, sometimes prepositions can be used as an indicator to the relations between human and object positions.

We want to exploit the connection between human-object positions in images and prepositions that link human, verb and object in language. Intuitively, if an action implies a strong positional relation between the human and the object, we expect to find specific, distinguishing prepositions in language. For example, in language you usually say "sit *on* chair", where the preposition *on* suggests a specific spatial relation between the human and the chair. When an action does not imply a strong positional relation, such as "play", we expect no specific prepositions.

***Links* in language models** To test this hypothesis, we use TypeDM (Baroni and Lenci, 2010), a distributional memory that has been built from large scale text corpora. This model contains weighted <word-link-word> tuples extracted from a dependency parse of corpora. The relations between words are characterized by their "link". Some of these links are prepositions that connect verbs and objects together. Examples of some tuples with word-link-word and their weights are provided in Figure 3.

Number of links and link entropy We want to determine whether there is any correlation between human-object relative positions in images and the associated prepositions from language models. To do this, we record two metrics: the number of links, where we count how many different links that connect verbs and objects in the language model; and the entropy of each action A^i verb-object pair (where the human is implicit) is $H(A^i)$ defined by: $H(A^i) = -\sum_{l_j \in L^i} p(l_j) \times \log p(l_j)$

where L^i is the set of all links that occur between verb and object of action i; $p(l_j)$ is the probability of the link l_j of the action A^i:



bicycle-n	by	ride-v	11.2994
bicycle-n	in	ride-v	6.7795
bicycle-n	of	ride-v	2.4167
bicycle-n	on	ride-v	278.4273
drum-n	against	play-v	3.5056
drum-n	behind	play-v	4.7656
drum-n	by	play-v	2.4393
drum-n	in	play-v	8.9440
drum-n	of	play-v	2.9940
drum-n	on	play-v	185.8888
drum-n	over	play-v	2.8841
accordion-n	on	play-v	174.7606
ant-n	over	hold-v	3.3807
apple-n	in	hold-v	0.3519
apple-n	on	hold-v	1.1309

Figure 3: Examples of word-link-word and their weights in the distributional memory

$$p(l_j) = \frac{weight(l_j)}{\sum_{l_k \in L_i} weight(l_k)} \tag{4}$$

where $weight(l_j)$ is the weight given by the TypeDM of link j in action i.

Generally, the entropy for each action allows seeing whether a link is predictable for a given pair of verb-object or not: when a link is predictable, the entropy is expected to be low (contain little information), which might correspond to the case that the position information will be useful in predicting actions and the other way around.

	Number of links	Entropy
Group 1 (position helps)	8	1.05
Group 2 (position doesn't help)	15.3	1.36

Table 6: Actions that can be disambiguated by positions (Group 1) vs. actions that cannot be disambiguated by positions (Group 2) and their links in the language model

Results The result shown in Table 6: for the first group (with position is better), the average number of links per relation (verb - object) is 8 and the average entropy is 1.05; the average number of links per relation for the second group is almost twice more, 15.3, and their average entropy is also higher, 1.36. It shows that verbs which have many different ways of linking to an object might not have a *representative* relative position between the person and object, hence more difficult to be classified based on their positions. Verbs that have less links to an object tend to have more *fixed* relative positions between persons and objects, hence it might be helpful to use position information in classification.

A qualitative analysis We further examine actions where this statement does not apply, i.e., actions with high number of links and high entropy but belong to group 1 (position information helps) and actions with low number of links and entropy belonging to group 2. For the first case, typical actions which have high number of links/entropy are: ride car, ride bus, ride train, pull cart, light lamp. The large number of links of these actions seem to come from relations which do not describe the human/object interaction itself. For example, the links associated with 'ride bus' do not all actually refer to 'ride a bus' but to ride another object in a position with respect to the bus: ride after bus, ride behind bus, ride before bus. These cause extra links which are not related to the action itself. Similarly, actions pull of/around/behind/below/on cart, there is another object which is moved to a specific position with regards to the cart.

For the second case, examples of typical actions with low links but for which positions information doesn't help are hold harmonica, wear diaper, hold ladle, spread cream, hold racket, apply lipstick. These actions are related to objects, for which their positions depends a lot on the human pose (e.g., hold something). These actions in the language model do not contain many links as we expected: the most possible link between hold, harmonica is *in*, which probably means hold harmonica *in* your hand.

Instead of looking at actions, we look into typical verbs where position information helps in classifying actions and verbs where position information doesn't help. For the first group, the most frequent verbs are: chop, cut, drink, feed, lean, sit on, sleep, look at, put on, shake, shoot, wash, catch. For the second group, the most frequent verbs are: clean, cook, lift, punch, sing, spray, spread. It can be observed that verbs related to some particular poses or relative positions between human and object are better with the position information (chop, drink, sit on, sleep), and verbs related to more various human poses and unspecific are not helped by the position information (cook, sing, spray, clean).

Generally, there is a relation between prepositions in language and the relative positions between human-object in images. Although this statement does not hold in every cases, for example when the prepositions refer to the positions between another action (e.g., ride) and that object (e.g., after a bike), this can be potentially solved by better NLP parsing and analyses of verb phrases. Furthermore, actions that cannot be disambiguated by positions are usually related to different human poses, while actions that have some particular human poses can be classified using position information.

5 Conclusion

In this paper, we have introduced the Trento Universal Human Object Interaction image dataset, TUHOI. This dataset contains more than two thousand human actions associated with 189 common objects in images. The main characteristics of this dataset are that it follows the actual human action distribution observed in images, it captures different ways of describing an action and it enables the study of how verbs are used differently with different objects in images. Additionally, we performed some preliminary experiments in which we show that action recognition can benefit from using position information. Finally, we showed that this position information is related to prepositions that can be extracted from a general language model.

References

Stephane Ayache and Georges Quenot. 2008. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, mar.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*.

Vincent Delaitre, Ivan Laptev, and Josef Sivic. 2010. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*. BMVA Press.

Vincent Delaitre, Josef Sivic, and Ivan Laptev. 2011. Learning person-object interactions for action recognition in still images. In *NIPS*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

Ali Farhadi, Mohsen Hejrati, Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences for images. In *ECCV*.

Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10), October.

Nazli Ikizler, Ramazan Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. 2008. Recognizing actions from still images. In *ICPR*, pages 1–4. IEEE.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander Berg, and Tamara Berg. 2011. Babytalk: Understanding and generating simple image descriptions. In *CVPR*.

Ivan Laptev. 2005. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September.

Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. 2013. Exploiting language models to recognize unseen actions. In *ICMR*.

Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *CVPR*.

Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition.

Bangpeng Yao and Fei-Fei Li. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, pages 9–16.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Li Fei-Fei. 2011. Action recognition by learning bases of action attributes and parts. In *ICCV*.