Audio-Visual Classification and Detection of Human Manipulation Actions

Alessandro Pieropan Giampiero Salvi Karl Pauwels Hedvig Kjellström

***Abstract*— Humans are able to merge information from multiple perceptional modalities and formulate a coherent representation of the world. Our thesis is that robots need to do the same in order to operate robustly and autonomously in an unstructured environment. It has also been shown in several fields that multiple sources of information can complement each other, overcoming the limitations of a single perceptual modality. Hence, in this paper we introduce a data set of actions that includes both visual data (RGB-D video and 6DOF object pose estimation) and acoustic data. We also propose a method for recognizing and segmenting actions from continuous audio-visual data. The proposed method is employed for extensive evaluation of the descriptive power of the two modalities, and we discuss how they can be used jointly to infer a coherent interpretation of the recorded action.**

I. INTRODUCTION

There has been tremendous effort in the robotics community to develop robots able to operate autonomously in unstructured environments. Robots should be able to perceive the world correctly, detect objects, observe and interact with humans, perform activities and understand if the desired outcome has been achieved [1].

Much effort has been spent on development of methods for visual perception and modeling of scenes. Many successful solutions have been proposed, considering various visual aspects such as object appearance, motion, human pose and affordances. Nevertheless, the use of visual features has some limitations. Firstly, an action has to be performed within the field of view of the observer in order to be perceived. Secondly, object detection and tracking are very sensitive to occlusions while performing activities. Finally, there are meaningful states induced by an action that are hard to detect just relying on visual perception, it is, e.g., very difficult to detect if a person has turned on an oven.

One approach to tackle these limitation is to use additional sources of information, e.g., audio. While this field was in the past more focused on solving problems such as source localization and noise filtering, lately, the focus is more on exploiting sound to extract semantic knowledge and perform scene understanding [2]. In many contexts, such information is very descriptive and can be helpful to understand a scene. Potentially it could compensate the limitations of visual perception (Fig. 2). However, an important question arises: how can different modalities contribute to formulate a coherent interpretation of the world an agent is observing, in a manner similar to how humans process and integrate multiple modalities [3]? Multimodality has up to now received relatively little focus in the robotics community, with some exceptions, e.g., [4], and the task of audio-visual action recognition has not, to our knowledge, been addressed in robotics.

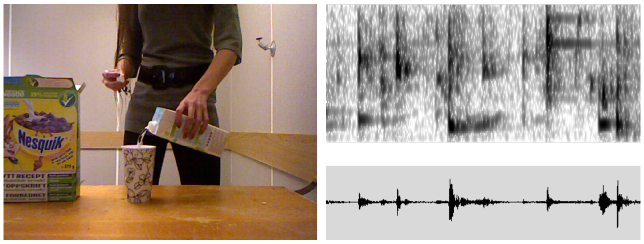
(a) RGB-D video (b) Audio

Fig. 1. An RGB-D video modality and an audio modality give complementary information about an observed human action. This is a motivation of why to use a recognition method that takes both modalities into account.

The main contribution of this paper is a *method for audio-visual recognition of human manipulation actions*. The visual features (Section VI-A) are the relative 3D orientations and positions of objects involved in the activity, while the audio is represented in terms of MFCC features [5] (Section VI-B). Objects are tracked in 3D from an RGB-D video stream using the method in [6] (Section IV). The audio and video cues are fused in an HMM framework (Section V).

An additional contribution of the paper is an *RGB-D-audio dataset of humans involved in the activity of making milk and cereals* (Section III). This composite activity consists of many sub-actions such as opening and closing boxes and bottles, and pouring milk and cereal. The recognition method described above was evaluated on this dataset (Section VII). The dataset is annotated with manually extracted segmentation points between sub-actions, and 3D positions and orientations of objects, obtained as described above. It also features 3D models of all objects involved. We release source code for administration of and access to the dataset.

II. RELATED WORK

Modeling and recognition of human activity from visual perception is an important problem tackled by the computer vision community, with applications in a wide variety of domains including health monitoring, visual surveillance, video search, human computer interaction and robot learning

This research has been supported by the EU through TOMSY, IST-FP7-Collaborative Project-270436, Marie Curie FP7-PEOPLE-2011-IEF-301144 and the Swedish Research Council (VR).

The GPU used for this research was donated by the NVIDIA Corporation.

AP and HK are with CVAP/CAS, KTH, Stockholm, Sweden, pieropan,hedvig@kth.se. GS is with TMH, KTH, Stockholm, Sweden, giampi@kth.se. KP is with the University of Granada, Spain, kpauwels@ugr.es.

from demonstration. It has been shown [7]–[15] that it is advantageous to represent human activity both in terms of human motion and the objects involved in the activity. Several approaches are widely used to tackle this problem. [16], [17] estimate the human body pose to recognize human activities. [18] estimates human motion trajectories to infer activities. [19] recognizes actions in videos by extracting low level spatio-temporal feature descriptors.

However in the context of robot learning from demonstration it is meaningful to focus the attention on descriptive features that encodes how objects are being used by a human [20], [21] to achieve a task. It is then possible that the robot can imitate the human and operate autonomously to accomplish given tasks. This has been confirmed in recent works [22]–[27] and it is supported by Gibson's affordance theory [28], which states that humans and animals recognize objects in terms of function, in a task oriented manner.

[29], [30] present a global representation of activities in terms of spatial relations between objects. The recognition of activities is determined on a set of pre-defined symbols which describe the relationships between segmented coherent regions. In our previous work we have show that is also possible to describe object functionality in terms of how they are being handled by a human [31]. Moreover, [23] simulates humans performing activities using objects and cluster them into functional classes. In [22], [32], [33], real humans are instead observed in interaction with a scene.

Visual perception supply part of the information to interpret the world. Much can be done from the auditory perspective. In [34] learning object affordances was extended to auditory perception. However, the acoustic information was in the form of linguistic descriptions of the scene. In this work, on the contrary, we use the sound generated directly by the interaction between objects as a result of human actions. This is similar to what has been done in [35]. However what we propose here is to go one step further, we want to evaluate different sources of information in order to find the limitations and see how they can be jointly used to overcome such limitations. This is in the spirit with a similar work [4] where language has been exploited to perform action recognition. in order and tackle the same problem but at a higher level using information or features coming from different perceptual input.

Most of the cited works tackles what is known in neuroscience as the binding problem [36], our brain formulate a coherent interpretation of the world from complex input merging multiple sources of information. Such a skill should be taken into account to develop autonomous robots.

III. Dataset

The data set we have collected includes observations of eight subjects fulfilling the task of making cereals. The actors are not instructed on how to perform the action and, therefore, there is substantial variation in the way they perform it. However, the action can generally be decomposed into 6 different sub-actions: open milk box, pour milk, close milk box, open cereal box, pour cereals and close cereal box. The variability can be observed in the order these sub actions are performed, the distribution of workload between left and right hand, the position of the objects, and so forth. Since the actors perform the action in a natural way the transitions between sub-actions are smooth, some sub actions can be performed in parallel and some may be missing (e.g. sometimes subjects leave the cereal box open at the end of the action.). In details the data set includes:

- calibrated RGB-D video recorded using a Kinect device with 30 Hz framerate and a resolution of 640x480. The time stamp of each frame has been saved so that it is possible to align audio and video correctly.
- 4 separate audio tracks using the Kinect microphone array sampled at 16 kHz with 32 bits depth.
- 25 3D object models, built from real images, saved in a standard format (.obj). The objects used in the experiments are 4 milk boxes, 2 cereal boxes and 5 cups. We release all objects as they can be used for other purposes such as grasp planning.
- object 6DOF pose trajectories for each recording.
- manual labels for each sequence including 6 different sub actions: *Open Milk Box*,*Pour Milk*,*Close Milk Box*, *Open Cereal Box*, *Pour Cereals*, *Close Cereal Box*. The labels are provided as a standard subtitle file for videos (.ass).
- python scripts to read the data, synchronized the sources and parse the labels are provided.

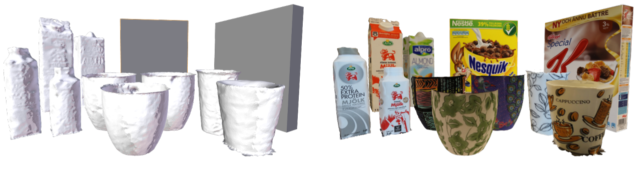
(a) Solid Models (b) Textures

Fig. 3. Samples of objects used in the dataset. (a) shows the 3D solid models of the objects, (b) shows the models rendered with textures.

IV. Pose Estimation

To estimate the objects' location and orientation over time we used a real-time method that relies on sparse keypoints for pose detection and dense motion and depth information for pose tracking [6]. This method can simultaneously track the pose of hundreds of arbitrarily-shaped rigid objects at 40 frames per second, with high accuracy and robustness. This is enabled through a tight integration of visual simulation and visual perception that relies heavily on Graphical Processing Units. A detailed 3D scene representation, consisting of the textured wireframe models from Fig.1, is constantly updated on the basis of the observed visual cues. Self-occlusions and occlusions between modeled objects are handled implicitly by rendering the scene through OpenGL. Pose detection runs in parallel with pose tracking, allowing for automatic pose initialization and recovery when tracking is lost.

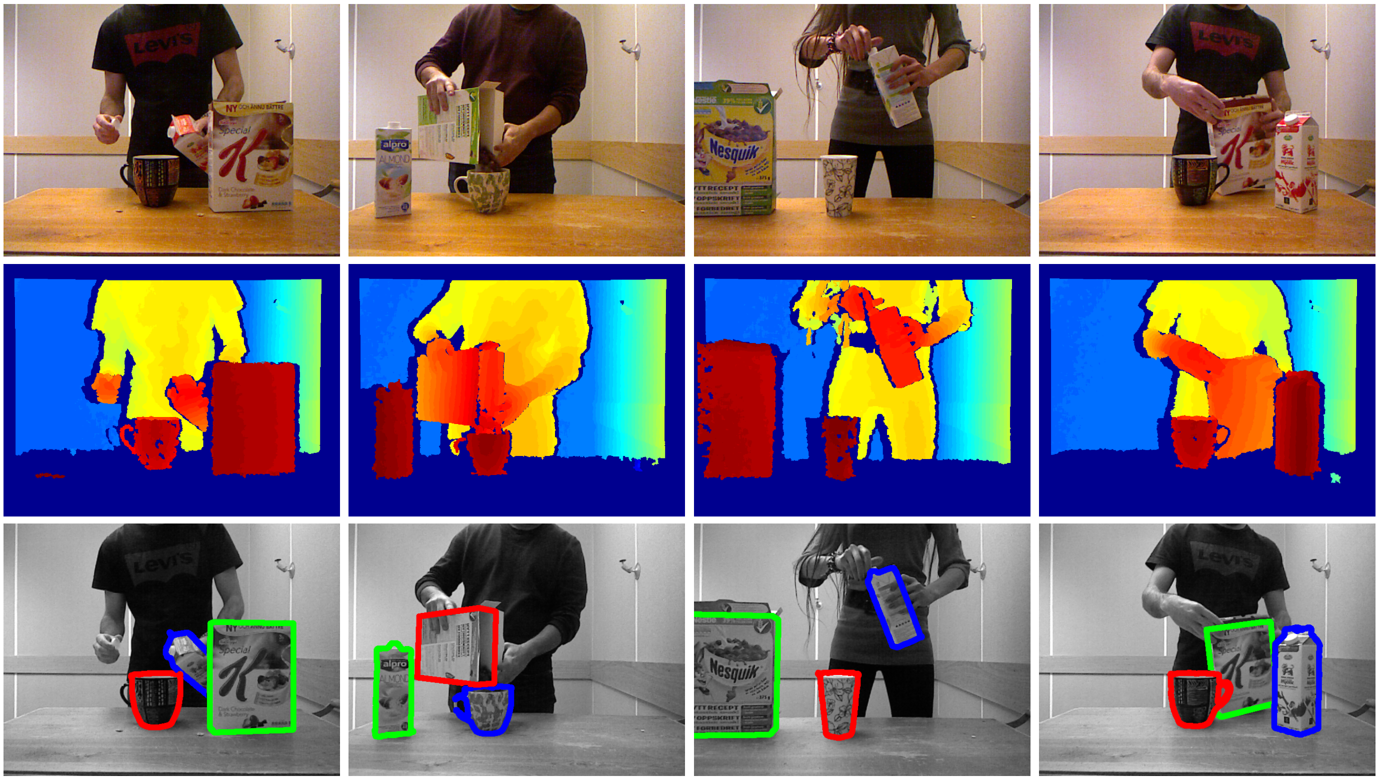

Fig. 2. Overview of the dataset. Activities are performed by different actors in a natural manner.

V. Model

Given a continuous data input recorded while performing an action (*making cereals*), our goal consists in recognizing the ongoing sub-actions and detecting them in the continuous stream of data. This is similar in a way to what is done in speech recognition to classify words and phrases, where our words are the sub-actions and a complete action corresponds to a phrase. Therefore we propose to model the recognition of each single sub action included in our data set using a left-to-right Hidden Markov Model (HMM) with continuous observation densities, which is a widely used method in speech recognition. Given the set of sub-actions L an instance of the lef-to-right model properly trained on a sub-action is then used to predict the correct label $l \in 1, \cdots, L$, where l are the labels present in the data set plus the label *garbage* used to collect uninformative parts. We describe the proposed model in the following Sec. V-A.

A. Recognition of isolated actions

Given an atomic action $l \in 1, \cdots, L$, a left-to-right HMM model $\lambda(l) = (A_l, B_l, \pi_l)$ is created, where A_l is the state transition model, B_l the state to observation model and π_l the state prior.

The model parameters are initialized in the following way: the state prior $\pi_l(i)$ is set 1 for the first state (i=1) and 0 otherwise. The transition matrix is initialized in such a way that from each state only two transitions are possible and equally likely: the self transition and the transition to the next state as in Eq. (1):

$$A_l(i,j) = \begin{cases} 0.5, & \text{if } j = i \text{ or } j = i+1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The observations are treated as a continuous observation density represented by a Gaussian Mixture as in the following equation:

$$B_l(j,O) = \sum_{m=1}^{M} c_{ij} N(O; \mu_{jm}, \Sigma_{jm}) \ , \quad 1 \leq j \leq N \tag{2}$$

Given such a model denoted as $\lambda_l = (A_l, B_l, \pi_l)$ is then possible to calculate the probability $P(O|\lambda_l)$ of a new sequence of observations $O = \{O_1 \cdots O_t\}$ by means of the forward algorithm. The best class given the observations is then:

$$L(O) = \arg\max P(O|\lambda_l), 1 < l < L. \tag{3}$$

B. Continuous action recognition

The modeling described in the previous section requires that the segmentation be available for the sub-actions to be recognized. This is not a valid assumption in a realistic scenario. We, therefore, introduce in this section a combined model that can be used for continuous action recognition and does not rely on sub-action pre-segmentation. The combined model is obtained by connecting the final state of each isolated model to the initial state of any other including itself. This so called *free loop* has been used in similar works done in gesture recognition (e.g. [37]), and is illustrated in Fig V-B. This allows to model any sequence of sub actions which may occur in our data.

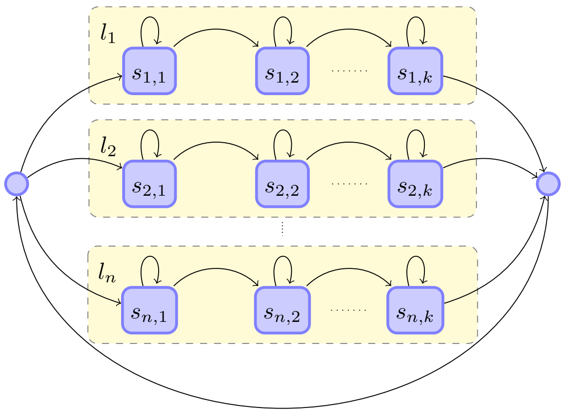

Fig. 4. Template of the composite hmm structure to recognize and segment activities.

The recognition in this case is performed by searching for the most likely path through the model by means of the Viterbi algorithm.

C. Feature fusion

One interesting problem that raises consists in how a multi-sensory input can be integrated to formulate a coherent interpretation of a scene. Ideally such fusion should mitigate the weaknesses of the single sources. The fusion can be performed at any level in the learning process. It can be done at a low level by considering the features coming from different sources as a single feature vector and train the model with that. It can be done at a middle level by designing a model such as coupled HMM as proposed by [38] in the context of generating lips movement from speech. It is also possible to design a more general Dynamic Bayesian Network (DBN) and defining the cross dependencies between the different sources of observation (e.g. [39]). It can be performed at high level having a different model for each single source and designing a voting system to infer the right interpretation, the non-Markovian esemble voting proposed in [35] is a valid mechanism. Every approach has its strengths and weaknesses, in this work we experiment only on the low level fusion by defining our feature vector as the concatenation of audio and video features: $O_t = [a_t, d_t], [a_t, d_t + \theta_t], [a_t, d_t + \theta_t + d_{\theta_t}]$, where the symbols are defined in Sec. VI. It is our intention to explore more possibilities as discussed in Sec. VIII.

VI. Features

A. Extraction of video features

In the spirit of recent works in activity recognition [29], [30] where human manipulation actions are represented in terms of the spatial relationships between the objects involved. The 6 DOF pose of all objects in the scene are estimated using the tracker described in Sec. IV.

The pairwise interaction between two objects i and j can be characterized by the pairwise distance over time, measured in terms of Euclidean distance,

$$d_{i,j,t} = \parallel C_{i,t} - C_{j,t} \parallel \qquad (4)$$

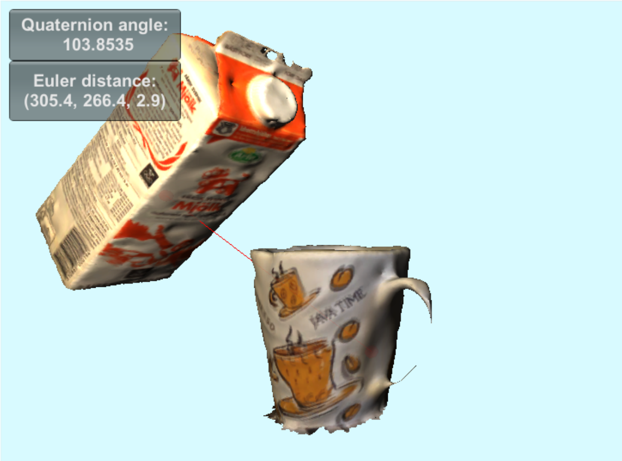

Fig. 5. Example of the rotation feature extracted between a pair of objects. The distance between quaternions measure the distance in all dimensions. The euler distance measures the distance between each dimension separately.

Euler angle distance, i.e., the difference between the Euler angles calculated from the rotation matrices at time step t,

$$d_{\theta_{i,j,t}} = \parallel (\alpha_{i,t}, \beta_{i,t}, \gamma_{i,t}) - (\alpha_{j,t}, \beta_{j,t}, \gamma_{j,t}) \parallel \qquad (5)$$

or the quaternion distance, calculated by extracting the quaternions from each object rotation matrix at time t and computing the angular value between each pair of object (Fig. VI-A) as follows:

$$\theta_{i,j,t} = 2 * \arccos\left(q_{i,t} \cdot q_{j,t}\right) \qquad (6)$$

1) Invariant Scene Description: The use of objects' spatial relationships as visual features raises an important problem: how can such features be an invariant descriptor of a scene? Given a set of unknown objects detected, their pairwise relationships cannot be used directly in a Hidden Markov Model because the order in which they are taken into account as observations may change the results. A possible solution is to represent the relationships as a scene graph proposed by [29] but, in their current implementation, this approach requires the manual definition of discrete labels, and it is furthermore sensitive to the observation point of view. We intend to keep the descriptive power of the continuous data and prefer to use a representation that is invariant to the point of view of the observer.

In the present method, each component sub-task is described exclusively by the pair of objects involved. As an example, the sub-action of pouring milk can be characterized by the tuple <milk container, cup>. In the training phase, each sub-action model is trained only with the features from the involved pair of objects (except for the garbage, no-action model, which is trained with all pairwise combinations of objects in each frame). During the testing phase the features for all pairs of objects are extracted from the data set, and each potential pair of objects are evaluated against all sub-action models using Eq. (3).

B. Extraction of audio features

The Kinect microphone array provides four audio channels that can be used for beam forming in order to reduce the effect of noise in unstructured environments. Because our recordings are in a quiet room, we decided to only use one channel for this experiments for simplicity. All four channels are, however, available for additional processing. For example, beam forming algorithms can be used to determine the orientation of the acoustic source relative to the device, thus adding an independent source of information.

From the selected audio channel, Mel Frequency Cepstral Coefficients (MFCC) [5] were extracted. This is one of the most robust and widely used sets of features in the field of audio-based feature extraction. MFCCs were designed primarily for speech recognition but there is a large body of work where they have been used to classify a broad set of different sound classes [35], [40].

The audio samples at 16 kHz are first grouped into overlapping windows (frames) of 25 ms length displaced at 8.3 ms intervals, resulting in a frame rate of 120 Hz. A Hamming window is then applied and the sound is pre-emphasized with coefficient 0.97. Then, up to 13 Mel Cepstral coefficients, including the zeroth coefficient, are computed using a Mel filterbank of 26 channels. First and second order time derivatives are computed on the 13 feature vectors resulting in a total of 39 coefficients per frame. In the experiments we denote the audio feature vector at any given time t as $a[1\cdots n]_t$ where $1 \leq n \leq 39$. We only use small subsets of these coefficients, see, e.g., Figure 6.

C. Feature processing

The features extracted from the two input sources have two issues: first they are recorded with a different frame rate due to the hardware specification, video frames are extracted at a frequency of 30 Hz while audio features are extracted with a frequency of 120 Hz, second the input source registration is not perfectly synchronized. To solve the first problem we generate from the video input intermediate synthetic data points by linearly interpolating the recorded features. This way the resulting features have the same frame rate. To solve the second issue we use the time stamp saved while recording the data set to find the time interval where both sources of information are recorded and discard all feature points that are not within that interval.

VII. Experimental Results

A. Isolated model training

The experiments on the isolated sub action recognition are performed in the following manner. We define the set L as the set of sub-actions composing the action of making cereal as explained in Sec. III plus a label *garbage* used to classify parts of the input data when nothing meaningful happens. The data set is split into training and test set. On those sets and for each label $l \in L$ audio and video features (Secs. VI-B and VI-A) are extracted. Each isolated model for the recognition of an action with label l, $\lambda(l) = (A_l, B_l, \pi_l)$, is initialized with 3 hidden states so that the transition matrix $A_l = 3 \times 3$ with the initial values defined by Eq. (1) and the value of $A_l(3,3) = 1$. The observation node is defined as a single multivariate Gaussian $B_l = N(x; \mu_j, \sigma_j)$ with $0 < j \leq 3$. The training features F_t are used to initialize μ_0 and σ_0 by computing the global mean and covariance of the features. The model is, then, trained using the training data, μ_0, $\sigma 0$ and a prior $\pi_l = (1,0,0)$. All experiments have been executed 10 times by randomizing the training and testing data set and the results are averaged out.

TABLE I

Classification of isolated actions performed using our model with the best parameter and feature setting. Average accuracy is 0.73%.

	OM	PM	CM	OB	PC	CB	N
Open Milk	**0.68**	0.02	0.17	0.03	0.00	0.02	0.07
Pour Milk	0.04	**0.89**	0.06	0.00	0.00	0.00	0.01
Close Milk	0.28	0.05	**0.59**	0.00	0.00	0.03	0.03
Open Box	0.03	0.00	0.01	**0.85**	0.01	0.08	0.00
Pour Cer.	0.00	0.06	0.00	0.04	**0.87**	0.01	0.00
Close Box	0.05	0.00	0.00	0.39	0.05	**0.47**	0.03
Null	0.07	0.04	0.03	0.00	0.00	0.08	**0.80**

B. Classification of segmented labels

Give the set of labeled actions (*open milk box, pour milk, close milk box, open cereal box, pour cereals, close cereal box, garbage*) we intend to explore the descriptive power of auditory and visual input.

1) Classification from audio feature: We perform our experiments to test the auditory accuracy using an incremental number of features, from 3 to 13. Fig.6 shows the overall results using subsets of features. From the experiments, it turns out that the optimal subset of features in these experimental conditions is $a[5]$. The confusion matrix using $a[5]$ is shown in Table I. The classification of pouring cereal and pouring milk is respectively 0.89 and 0.87, while the detection of opening and closing is lower. We expected this behavior as the sounds to open and close the same object is similar. The confusion matrix confirms our expectation.

It can be noticed that the features allow a rich granularity for recognition making possible not only to detect the *pouring* action but also to discern with high precision pouring milk from pouring cereals. This is a very relevant piece of information when a robot needs to understand the fulfillment of an action, it is not sufficient to detect that the *pouring* actions has been performed twice to infer that cereals has been done (i.e. a human could have poured more milk twice). Moreover, it is also possible to classify opening and closing of the corresponding containers, such information is very difficult to extract from a visual input as the milk box top is very small and the cereal box needs a model that can reason about deformable objects or complex objects with some constrained joints.

2) Classification from video feature: The classification of actions may be trivial since the object class is known in our data set. However such an approach will not scale well in

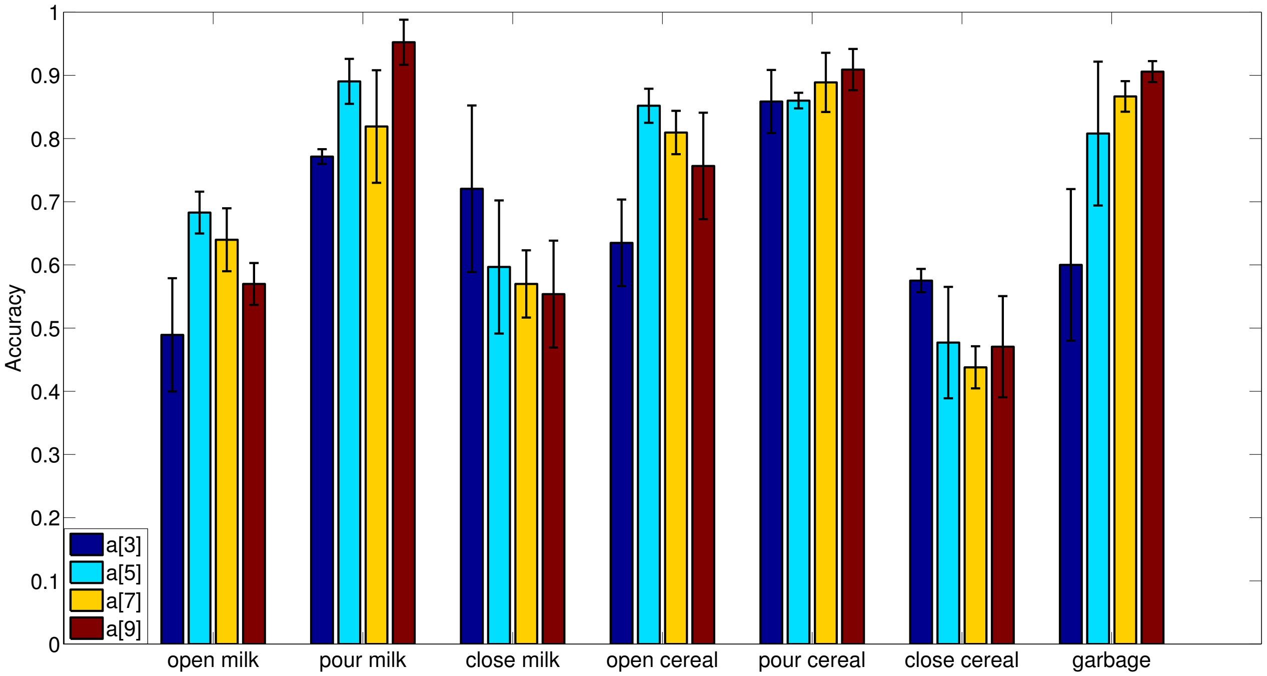

Fig. 6. Test performed using an incremental number of audio features $a_{1\cdots n}$. The image show sample results where $n = 3,5,7,9$. Upon inspection $n = 5$ has been chosen as optimal feature vector to perform further recognition tasks.

real scenarios as it requires all possible instances of object classes. Moreover, the mapping between visual appearance and object functionality is not necessarily one-to-one [23], [31]. We focus here on how objects can be characterized in terms of *how they normally interact with each other*. We believe that such representation can be more flexible doing recognition in a unstructured environment with previously unseen object instances.

The experimental results in Table II show that *pouring milk* can be recognized with an average accuracy of 0.72 and *pouring cereals* with an accuracy of 0.54 in the best case, obtained with the distance feature d. However, the features are not very accurate in the classification of the other labels. This is expected as there is a high variation in the way this activity has been performed by the different actors. Moreover, actions like *open, close* often involved only one object, therefore a pairwise feature between objects could not capture them well.

TABLE II

CLASSIFICATION ACCURACY ON ALL ACTION LABELS USING VIDEO FEATURES.

	OM	PM	CM	OB	PC	CB	N
d	0.18	**0.72**	0.04	0.15	**0.54**	0.27	0.20
d+θ	0.15	**0.53**	0.31	0.17	**0.46**	0.60	0.23
d+d_θ	0.28	**0.41**	0.22	0.19	**0.49**	0.13	0.29
d+θ+d_θ	0.40	**0.31**	0.19	0.21	**0.30**	0.22	0.38

As the features are not very descriptive for activities such as *open, close* we perform a second experiment where the features are used to train 3 classes of labels: *pour milk, pour cereals, garbage*. All the features extracted while performing the other actions are then treated as noise. Clearly from the results shown in Table III it can be noticed that the proposed model improves in detecting the activities meaning that the features capture the undergoing action.

On a first look the results in Table III, it can be noticed that the accuracy of the feature d is 0.80 while $d+\theta$ is 0.85 for *pouring milk*. This behavior is justified by the high variance in the way actions are performed by the actors. As an example some actors place the milk box close to the cup after pouring the milk. The *pour* action looks then similar to an *idle* state in those experiments. This behavior is reflected in the results where the classification based purely on distance has high misclassification rate for the *garbage* class. Another example can be when an actor opens a milk box or a cereal box while holding it tilted immediately before pouring. This results in higher misclassification for the *pour* class when using θ or d_{theta} features. However, the classification improves for the *garbage* class.

TABLE III

CLASSIFICATION ACCURACY TRAINING MODELS FOR POURING ACTIVITIES AND REGARDING ANYTHING ELSE AS GARBAGE.

	PM	PC	N
d	0.89	0.80	0.28
d+θ	0.72	0.85	0.37
d+d_θ	0.66	0.60	0.63
d+θ+d_θ	0.66	0.71	0.69

3) Classification from audio and video features: As discussed in Sec. V-C we perform classification experiments to see if we can achieve higher accuracy by merging the differen

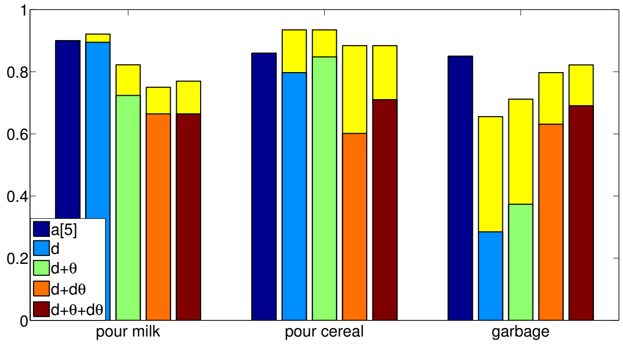

Fig. 7. Test performed using audio and video features together. Bar in yellow show the increment in recognition using a concatenation of audio and video features $[a,d] \cdots [a,d,\theta,d_\theta]$ with respect to the classification just with the corresponding video features $d \cdots [d,\theta,d_\theta]$.

input features. This is done by training an isolated model for each label with a feature vector having audio and video features concatenated. It can be seen in Fig. 7 that the integration of audio in the classification consistently improves the classification with respect to the pure classification based on visual feature. The average improvement is 0.08 for the *pour milk* class, 0.17 for the *pour cereal* class and 0.25 for the *garbage*. As shown in the middle block of columns, the joint use of audio with any combination of video features outperform the classification of *pour cereal* done only with audio features. This indicates that the visual characteristics of this action (shaking motion) are complementary to the audio characteristics (rattling sound) and interact to help raise the classification performance.

C. Classification and segmentation of activities

Our last experiment uses the features to recognize and segment sub-actions given a continuous input data. We decide to use audio features for this experiment as it has good accuracy on all the label classes in the Data Set. The model used for this task is explained in Sec. V-B. Observations are extracted from each video in the test set and the best path trough the model is computed running the Viterbi algorithm. The generated sequence of states is then compared with the ground truth and the accuracy is calculated by counting the exact number of frames in which the ground truth and the sequence produced have the same label. The confusion matrix in Table IV show the results. It can be notice that while there is a high accuracy for the sub-actions of pour milk and pour cereals, there is missclassification between the respective *open* and *close* sub-actions. This happens because the sounds is similar and we define no constrain on the transitions between different states. We expect to improve the accuracy of the sub-action by adding prior knowledge on the structure of action in form of constraints in the transition matrix (i.e. open milk box cannot happen after pour milk as the box is already open.). The high miss classification rate of the garbage class is justified by the fact that each labeled sub-actions does not start and end exactly with the corresponding sound and therefore there is always noise at the beginning and end confusing the system.

TABLE IV

DETECTION AND SEGMENTATION OF ACTIONS GIVEN A CONTINUOUS INPUT DATA. ACCURACY IS CALCULATED CONSIDERING ALL TIME STEPS COMPARING THE RESULT PRODUCED BY THE VITERBI ALGORITHM AND THE MANUAL LABEL.

	OM	PM	CM	OB	PC	CB	N
Open Milk	**0,46**	0,19	0,17	0,05	0,00	0,04	0,09
Pour Milk	0.05	**0.69**	0.10	0.03	0.05	0.02	0.05
Close Milk	0.27	0.23	**0.35**	0.03	0.00	0.04	0.07
Open Box	0.05	0.00	0.02	**0.63**	0.09	0.18	0.00
Pour Cer.	0.01	0.02	0.00	0.09	**0.83**	0.04	0.00
Close Box	0.05	0.05	0.04	0.34	0.10	**0.38**	0.02
Null	0.13	0.29	0.06	0.10	0.08	0.06	**0.27**

VIII. CONCLUSIONS

We present a data set as a resource for studies in the fields of action recognition, object detection and multi-modal fusion. The data set contains an extended set of examples of *making cereals* action executed by 8 actors. It will be released publicly and it will be maintained and improved extracting more features such as hand tracking and recording more indoor kitchen actions. We also quantitative experiment studies in the context of action recognition proposing two models for the tasks of sub-action recognition from pre-segmented data and action segmentation from continuous input. Experiments show that the sub-actions can be classified with good accuracy and the fusion of multiple modalities can overcome the limitation of the single sources.

A. Future work

Experiments proved that the visual feature extracted are good representation of actions that involves multiple objects. However they are not descriptive enough to perform recognition of action such as *open* or *close* of objects. We intend to extract more visual feature to include in the data set such as hand pose. It is our intention to find a high level visual feature descriptor that is invariant from the position of actors and objects and that can scale well as the complexity of the scene increases. Finally we want to explore more the topic of multi-sensory fusion as we believe it is essential to achieve autonomous robots.

REFERENCES

[1] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory motor coordination to imitation. *IEEE TRO*, 24(1):15–26, 2008.

[2] R. F. Lyon. Machine hearing: An emerging field. *Signal Processing Magazine*, 27(5):131–139, 2010.

[3] S. Zmigrod and B. Hommel. Feature integration across multimodal perception and action: a review. *Multisensory research*, 26:143–157, 2013.

[4] C. L. Teo, Y. Yang, H. Daumé III, C. Fermüller, and Y. Aloimonos. Towards a watson that sees: Language-guided action recognition for robots. In *ICRA*, pages 374–381. IEEE, 2012.

[5] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman. Speaker indentification using mel frequency cepstral coefficients. In *Conference on Electrical and Computer Engineering*, 2004.

[6] K. Pauwels, L. Rubio, J. Diaz Alonso, and E. Ros. Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues. In *CVPR*, 2013.

[7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR*, 2010.

[8] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009.

[9] H. Kjellström. Contextual action recognition. In T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors, *Guide to Visual Analysis of Humans: Looking at People*, chapter 18. Springer, 2011.

[10] H. Kjellström, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 115(1):81–90, 2011.

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

[12] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *ICCV*, 1999.

[13] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *ICCV*, 2005.

[14] M. Veloso, F. von Hundelshausen, and P. E. Rybski. Learning visual object definitions by observing human activities. In *IEEE-RAS International Conference on Humanoid Robots*, 2005.

[15] 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.

[16] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation, 2010.

[17] Salman Aslam, Christopher F. Barnes, and Aaron F. Bobick. Video action recognition using residual vector quantization and hidden markov models. In *IPCV*, pages 659–666, 2010.

[18] Aaron F. Bobick, James W. Davis, Ieee Computer Society, and Ieee Computer Society. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[19] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003.

[20] A. Billard, S. Calinon, R. Dillman, and S. Schaal. Robot programming by demonstration. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, chapter 59. Springer, 2008.

[21] T. Lang and M. Toussaint. Planning with noisy probabilistic relational rules. *Journal of Artificial Intelligence Research*, 39:1–49, 2010.

[22] J. Gall, A. Fossati, and L. van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[23] H. Grabner, J. Gall, and L. van Gool. What makes a chair a chair? In *CVPR*, 2011.

[24] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *CVIU*, 62(2):164–176, 1995.

[25] L. Stark and K. Bowyer. *Generic Object Recognition using Form and Function*. World Scientific Series in Machine Perception and Artificial Intelligence – Vol. 10, 1996.

[26] M. A. Sutton and L. Stark. Function-based reasoning for goal-oriented image segmentation. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, chapter 59. Springer, 2008.

[27] M. W. Turek, A. Hoggs, and R. Collins. Unsupervised learning of functional categories in video scenes. In *ECCV*, 2010.

[28] J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.

[29] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen. Categorizing object-action relations from semantic scene graphs. In *ICRA*, 2010.

[30] G. Luo, N. Bergström, C. H. Ek, and D. Kragic. Representing actions with kernels. In *IROS*, 2011.

[31] A. Pieropan, C. H. Ek, and H. Kjellström. Functional object descriptors for human activity modeling. In *ICRA*, 2013.

[32] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3D scene geometry to human workspace. In *CVPR*, 2011.

[33] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *ICRA*, 2012.

[34] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor. Language bootstrapping: Learning word meanings from perception-action association. *IEEE SMC-B*, 42(3):660–671, 2012.

[35] J. Stork, L. Spinello, J. Silva, and K. O. Arras. Audio-based human activity recognition using non-markovian ensemble voting. In *RO-MAN*, pages 509–514, 2012.

[36] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6:171–8, 1996.

[37] G. Saponaro, G. Salvi, and A. Bernardino. Robot anticipation of human intentions through continuous gesture recognition. In *International Workshop on Collaborative Robots and Human Robot Interaction*, 2013.

[38] G. Englebienne, T. F.Cootes, and M. Rattray. A probabilistic model for generating realistic lip movements from speech. In *NIPS*, 2007.

[39] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein. Gesture-based dynamic bayesian network for noise robust speech recognition. In *ICASSP*, pages 5172–5175. IEEE, 2011.

[40] M. Mckinney and J. Breebaart. Features for audio and music classification. In *International Symposium on Music Information Retrieval*, pages 151–158, 2003.