# Combining Contextual and Modal Action Information into a Weighted Multikernel SVM for Human Action Recognition

Jordi Bautista-Ballester[1,2], Jaume Vergés-Llahí[1] and Domenec Puig[2]
[1]*ATEKNEA Solutions,Víctor Pradera, 45, 08940, Cornellà de Llobregat, Spain*
[2]*Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007, Tarragona, Spain*

Keywords: Multimodal Learning, Action Recognition, Bag of Visual Words, Multikernel Support Vector Machines.

Abstract: Understanding human activities is one of the most challenging modern topics for robots. Either for imitation or anticipation, robots must recognize which action is performed by humans when they operate in a human environment. Action classification using a Bag of Words (BoW) representation has shown computational simplicity and good performance, but the increasing number of categories, including actions with high confusion, and the addition, especially in human robot interactions, of significant contextual and multimodal information has led most authors to focus their efforts on the combination of image descriptors. In this field, we propose the Contextual and Modal MultiKernel Learning Support Vector Machine (CMMKL-SVM). We introduce contextual information -objects directly related to the performed action by calculating the codebook from a set of points belonging to objects- and multimodal information -features from depth and 3D images resulting in a set of two extra modalities of information in addition to RGB images-. We code the action videos using a BoW representation with both contextual and modal information and introduce them to the optimal SVM kernel as a linear combination of single kernels weighted by learning. Experiments have been carried out on two action databases, CAD-120 and HMDB. The upturn achieved with our approach attained the same results for high constrained databases with respect to other similar approaches of the state of the art and it is much better as much realistic is the database, reaching a performance improvement of 14.27% for HMDB.

## 1 INTRODUCTION

Analyzing video content has become critical in human robot interactions, where a robot must make a decision considering the information extracted from sensors such as cameras or lasers. In this context, our research focuses on the recognition of action in videos containing multimodal and contextual information about the means by which an action is carried out. Some public databases are conformed by a set of RGB videos where scenes and parameters such as illumination, focus, distance, and viewpoints are mostly controlled, and few information exits about the tools and objects that were involved in the action. In robotic contexts, it is usual to have multimodal information, provided by distance laser sensors or by 3D cameras such as Kinect.

CAD120 database (Koppula et al., 2013) is recorded with a high controlled environment, which is ideal for human-robot interactions, although it includes both contextual and multimodal information. This database contains 10 high level actions performed by 4 different subjects which in total corresponds to 124 manually annotated videos. However, in order to go beyond the current state of the art in action recognition topic for real videos, more realistic databases have been increasingly employed, including videos that stage more realistic actions.

HMDB (Kuehne et al., 2011), is one of the largest action video database to-date with 51 action categories, which in total contains 6849 manually annotated clips extracted from a variety of sources ranging from digitized movies to YouTube videos. This database has been created to evaluate the performance of computer vision systems for action recognition and explore the robustness of these methods under various conditions such as cluttered backgrounds, fast irregular motions, occlusions and camera motion. In this database, actions with contextually connected objects can be found, although no multimodal recording is available.

Specifically, we select from the HMDB database a subset of actions that are performed employing a tool or object. This contextual information allows the computer to discriminate apparently similar actions

Bautista-Ballester, J., Vergés-Llahí, J. and Puig, D.
Combining Contextual and Modal Action Information into a Weighted Multikernel SVM for Human Action Recognition.
DOI: 10.5220/0005669002990307
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 299-307
ISBN: 978-989-758-175-5

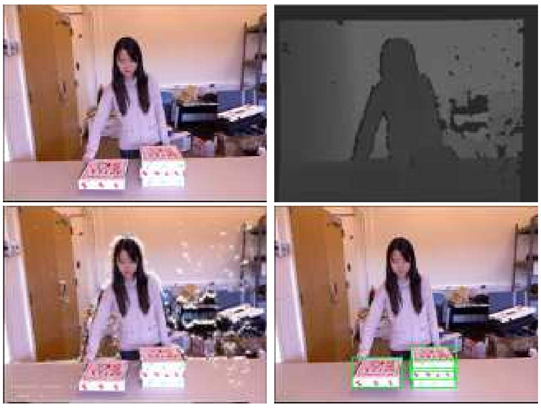


Figure 1: Multimodal database CAD120 with RGB (top left), Depth map (top right), 3D map (bottom left), object context (bottom right).

such as the case of shooting a gun or a bow. The biggest difference among these similar actions lies in the tool employed to carry out the action.

In this paper, we detail how these different sources of information -depth, objects- can be combined in a richer description of human actions that permits higher recognition rates. In order to increase the robustness of the recognition of actions in more challenging situations, we propose to weight different sources of information relevant to discriminate actions, namely, the spatio-temporal features that describe motion by RGB and depth modes, and the contextual information that explains how an action is carried out by object features. The Fig.1 shows sample images from CAD120 databes, representing the same frame of a video as four different sources of information: RGB, depth and precomputed 3D images and the objects detected in this frame.

The main contribution of this paper is the fusion and discrimination of new information sources for performed actions with a recognition structure that weights the addition of new information using a multichannel SVM. The use of the multichannel SVM has previously proven very successful in action recognition (Wang et al., 2013; Bilinski and Corvee, 2013). Thus, we take advantage of this structure in two ways: firstly, by adding data that is strictly not a descriptor of motion but modal or contextual information obtained by segmenting the region where the action takes place in three space dimensions and describing the tool employed in the action, which is a new way of using multichannel SVM. Secondly, by weighting the channels with a multikernel learning approach, determining which channel has more non-redundant information.

The paper is organized as follows. First, a review of the previous work is done in Section 2. In Section 3 our proposed approach CMMKL-SVM for combining multimodal and contextual information is detailed. The experimental setup including both databases used to evaluate our method is explained in Section 4. In Section 5 our experimental results over the two databases and comparisons with the state of the art are presented. Finally, in Section 6 the advantages of the proposed methodology is discussed and the paper concludes with future directions of the work.

## 2 RELATED WORK

Local space-time features (Laptev, 2005) have been shown to be successful for general action recognition because they avoid non-trivial pre-processing steps, such as tracking and segmentation, and provide descriptors invariant to illumination and camera motion. In particular, HOG3D (Kläser et al., 2008) has proven to outperform most of the descriptors of the same kind.

Experimenting in robotic environments, contextual and multimodal information have been considered in action recognition frameworks. Works in (Pieropan et al., 2014; Tsai et al., 2013) fuse information into two different stages with respect to the training, that is, before and after it respectively. In (Snoek et al., 2005) the authors studied the different methods of descriptor fusion and classified them into *early* fusion and *late* fusion approaches. The former consists of a fusion before the training step, while the latter is a fusion afterwards. In this context, (Ikizler-Cinbis and Sclaroff, 2010) combines six different visual descriptors for three different contextual information types, namely, *people* (HOF and HOG3D), *objects* (HOF and HOG), and *scene* (GIST and color histograms) by using a multiple MIL approach, which is a concatenation of bag representations and classified with an L2-Regularized Linear SVM. In (Bilinski and Corvee, 2013), a multichannel $\chi^2$ kernel SVM is used for the combination of a set of descriptors. Similarly, the work in (Wang et al., 2013) computes dense trajectories and their descriptors to finally combine them using an averaged multichannel SVM.

Considering multikernel learning (MKL) as an early fusion approach, it was first proposed in (Lanckriet et al., 2004). MKL approaches focus their efforts on how to improve the classification accuracy by exploiting different formulations and how to improve learning efficiency by exploiting different optimization techniques. The authors in (Bucak et al., 2014) showed that conflicting statements exist which are largely due to the variations in the experimental conditions. In this work it is also stated that while some studies reported that averaging kernels (same weight for each kernel) is outperformed by MKL (Bucak et al., 2010), others conclude the opposite (Gehler and Nowozin, 2009). Linear combinations do not have to deal with non-convex optimization problems which would lead them to poor computational efficiency and suboptimal performance. That is the reason why most of the authors prefer them instead of non linear combinations.

Traditional kernel combination learning approaches based on the MKL wrapper SimpleMKL (Rakotomamonjy et al., 2008) are mainly focused on the usage of the same training data, making use of linear, polynomial or RBF kernels. This fact is in contrast to recently published works on the multichannel approach in (Wang et al., 2011; Wang and Schmid, 2013), which combine different training data by kernel average.

In our work, unlike the aforementioned state-of-the-art methods, we consider depth, 3D information, and image descriptors of the objects used in the actions by means of a BoW-based action recognition approach. To this end, we first detect the set of points belonging to the object as in (Bautista-Ballester et al., 2014). Then, we compute codebooks for each video mode and context descriptors. Finally, we combine the three sources of information, motion, depth and objects, by weighting a multikernel SVM using CMMKL-SVM. Experimental results show that this procedure improves the recognition rate of actions.

# 3 METHOD

Our main goal in this work is to autonomously fuse and select the best information related to the action performed by means of a BoW-based representation of the action. In Section 3.1 which multimodal information we make use of in order to improve the action recognition performance is explained. In Section 3.2 how we utilize contextual information, i.e. objects, is described, firstly labeling its bounding boxes and, then, filtering the points used to construct codebooks. In Section 3.3 how we fuse and discriminate all the informational channels is explained.

## 3.1 RGB, Depth and 3D Multimodal

RGB images are usually provided by a single camera mounted in the body of the robot or in a fixed place in the space. That imposes the limitation of a single view of the performed action. Databases exist which consider the possibility of a multiple viewpoint, introducing more variability to the information captured. That would be the case if different robots were analyzing the same action simultaneously in different positions, but we consider human-robot interactions that involve just one robot. Hence, we test our algorithm over a database which provides depth maps, i.e. CAD120.

We make use of depth information in two ways: first, extracting descriptors as done with the RGB video sequences. We have, then, a set of descriptors such as trajectories, HOG, HOF, MBH for RGB and Depth. Depth sequences allow to differentiate elements in the scene like background and objects over planes different from the one in which the action takes place. Second, generating a RGB-D sequence in which we can extract 3D spatial descriptors, such as FPFH. 3D sequences provide 3D spatial information combined in one descriptor. In the end, RGB, Depth and 3D descriptors generate independent codebooks.

## 3.2 Object Detection and Tracking

In order to detect and track the objects in video sequences, we follow the work of (Bautista-Ballester et al., 2014). This method has been demonstrated to be successful in the addition of contextual information concerning objects related to the action. Considering that each video contains one action, we detect the objects that are employed in the performance of this action. We make use of the matching procedure based on the epipolar geometry, that computes the Fundamental Matrix between two consecutive frames and extracts the bounding boxes for each object in each frame. The result of this procedure is a set of bounding boxes that enclose the objects used in each action for each frame in the video ensuring high accuracy around the area that limits the objects. We also limit the computational burden by keeping a maximum of 100k points belonging to objects applying bounding box labels when creating the codebook like in (Bautista-Ballester et al., 2014).

## 3.3 CMMKL-SVM

Visual features extracted from a RGB video can represent a wide variety of information, such as scene (e.g., GIST), motion (e.g., HOF, MBH) or even just color (color histograms). In our approach we include extra features, such as depth and 3D scene information (e.g. FPFH (Rusu, 2009)), and object related information (e.g. (Bautista-Ballester et al., 2014)). To classify actions using all these features, the information must be fused in an appropriate way. According to the moment of the combination, (Snoek et al., 2005) proposed a classification of the fusion schemes in *early* or *late* fusion. Multikernel approaches use early fusion since the combination is done before the training.

The works of (Wang et al., 2013)(Bautista-Ballester et al., 2014) use a linear combination of different kernels, calculated from a set of codebooks generated with different descriptors. A SVM with a $\chi^2$ kernel for classification is used,

$$\chi^2(h_i,h_j) = \frac{1}{2}\sum_{k=1}^{n}\left(\frac{(h_i(k)-h_j(k))^2}{h_i(k)+h_j(k)}\right) \tag{1}$$

ensuring that the kernel matrices are strictly positive definite. They fuse different descriptors by summing up the corresponding kernel matrices, normalized by the average distance $A^c$ of $\chi^2$ distances between the training samples for the $c$-th channel. No kernel weighting is done, so no kernel is more discriminative than the others.

In our approach, given the base kernels

$$K_c(h_i,h_j) = exp(-\frac{1}{A^c}\chi^2(h_i^c,h_j^c)) \tag{2}$$

the optimal kernel of a certain descriptor is approximated as

$$K_{opt} = \sum_c d_c K_c \tag{3}$$

where $d_c$ is the kernel weight for $c$-th channel. Each $K_c$ represents the precoded $c$-th information referred to the action.

The optimization is carried out within a SVM framework that achieves the best classification on the training set subject to a regularization scheme. In this formulation, the objective function is near identical to the standard $L_1$ C-SVM objective function. The regularization prevents the weights from becoming too large, although this could be achieved by requiring that the weights sum up to the unit but also restricting the search space.

$$\begin{aligned} \underset{w,d,\xi}{\text{minimize}} \quad & \frac{1}{2}w^t w + C1^t\xi + \sigma^t d \\ \text{subject to} \quad & y_i(w^t K_c + b) \geq 1 - \xi_i \\ & \xi \geq 0, d \geq 0, Ad \geq p \end{aligned} \tag{4}$$

The constraints are also similar to the standard SVM formulation, with the addition of two constraints. First, $d \geq 0$, which ensures that the weights can be interpreted and also leads to a much more efficient optimization problem. Second, $Ad \geq p$, with some restrictions, that allow us to encode prior knowledge about the problem.

In order to tackle large scale problems involving hundreds of kernels, we adopt the minimax optimization strategy and solve the problem by using projected gradient descent, taking care to ensure that the constraints $dn+1 \geq 0$ and $Adn+1 \geq p$ are satisfied. This algorithm proceeds in two stages. In the first one, weights $d_c$ are maximized and support vectors (SV) obtained. In the second stage, objective function is minimised by projected gradient descent. The two stages are repeated until convergence or a maximum of the number of iterations is reached, at which point the weights $d$ and $SV's$ are obtained.

## 4 EXPERIMENTAL SETUP

In this section, modal selection and object detection and tracking are considered in detail. Afterwards, we introduce the encoding framework based on BoW. Finally, the databases and their experimental setups are exposed.

### 4.1 Extracting Contextual Information: Objects

The points used to identify and track the objects are a mixture of RGB points obtained using Harris corner detector and features computed applying SURF. We use a threshold between 0,04 and 0,1 for Harris detector and a maximum number of 1000 points for SURF. This ensures enough quantity of points with enough quality belonging to the object, even in the case that the object appearing in the video sequence is relatively small, like a ball or a sword. For the matching, we select the strongest 1% of matches, which is restrictive but ensures better point correspondences. These considerations refer mainly to HMDB database, which is more realistic than CAD120. Object detection and tracking for CAD120 are more accurate due to their highly controlled conditions.

### 4.2 Extracting Multimodal Information: RGB, Depth and 3D

We select three informational modes taking advantage of the RGB-D videos, forming the set with RGB, depth and 3D videos.

First, for each point in RGB and Depth videos we compute different descriptors, HOG3D, trajectories, HOG, HOF, MBH. In the case of HOG3D descriptors, we set the parameters optimized for KTH database as described in (Kläser et al., 2008), which have demonstrated a good performance not only for the KTH set, resulting in 1008 dimensions in total. In the case of trajectories, HOG, HOF, and MBH, we follow the work of (Wang et al., 2013) and set the parameters likewise. The dimensions of these descriptors are, respectively, 30 (trajectories), 96 (HOG), 108 (HOF) and 192 (MBH), which are significantly smaller than these of HOG3D. We set same parameter values for both, RGB and Depth videos.

Second, we consider the FPFH descriptor (Rusu, 2009) of the 3D Point Cloud Library. We configure the descriptor length to FPFHSignature33, that creates a 33 dimension descriptor. We set the FPFH radius search to 100 in order to ensure enough valid descriptors.

## 4.3 Encoding using BoW

We use the BoW approach to encode frames. First, we make use of STIP points following the work in (Laptev, 2005). We compute different descriptors for each point in RGB videos, Depth videos and 3D videos. We train a codebook for each descriptor type using a maximum of 100k randomly sampled features. For the object kernel, we ensure the object point selection using the method described in (Bautista-Ballester et al., 2014).

Afterwards, we group the points employing the k-Means clustering algorithm with a maximum of 5 iterations which ensures enough convergence. In order to compare results with (Bautista-Ballester et al., 2014), the size of the codebook is set to 500 words, avoiding over-learning, despite the fact that the larger the number of clusters employed, the better the performance is. Finally, a SVM with an exponential $\chi^2$ kernel is used for classification, using a 10 fold cross-validation method with the one-against-all approach. For all the experiments we employ the default parameter values in the LibSVM library (Chang and Lin, 2011).

## 4.4 Multikernel Selection

We perform a CMMKL-SVM for classification that uses the default parameters in (Vedaldi et al., 2009). We precalculate each kernel based on image coders (objects, 3D, Depth, RGB descriptors) and perform a train in order to obtain the best combination of weights.

In the comparison step, we also perform a uniformly weighted combination by summing their kernel matrices and normalizing the result by the average distance as in (Bautista-Ballester et al., 2014).

## 4.5 Databases

We test our model with two different databases, CAD-120 (Koppula et al., 2013) and HMDB (Kuehne et al., 2011). CAD-120 contains objects that involve actions in a highly controlled environment and multimodal information such as RGB and depth videos. HMDB is a more challenging and realistic one, where objects used in actions are present. Although no 3D information exists, we use this dataset to test our approach and compare the results to the state-of-the-art results. Sample frames for each database are shown in Fig. 2, in which three actions from the whole collection are represented for both databases.

### 4.5.1 CAD-120 Database (Koppula et al., 2013)

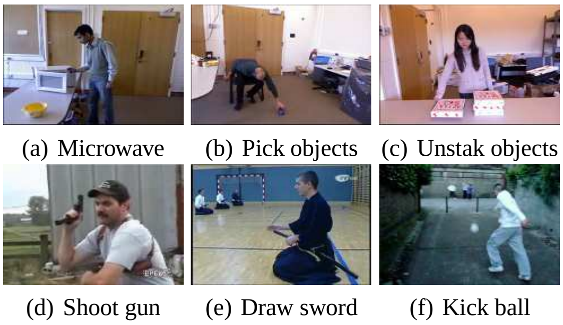
The CAD-120 database contains 124 RGB-D videos of 4 different subjects performing 10 high-level actions. Each action is performed three times with different objects. It contains a total of 61585 3D video frames. The actions have a long sequence of subactivities which might be considered in future work.

The 10 high-level actions performed are *arranging objects*, *cleaning objects*, *having meal*, *making cereal*, *microwaving food*, *picking objects*, *stacking objects*, *taking food*, *taking medicine* and *unstaking objects*.

### 4.5.2 HMDB Database (Kuehne et al., 2011)

The HMDB database consists of 51 actions from a total of 6,849 videos collected from a variety of sources ranging from digitized movies to YouTube videos. The action categories are grouped in five types: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction, and body movements for human interaction.

In order to obtain comparable results and considering that we need actions where an object is used, we do not follow the original splits proposed by (Kuehne



(a) Microwave (b) Pick objects (c) Unstak objects

(d) Shoot gun (e) Draw sword (f) Kick ball

Figure 2: Example frames from the CAD120 database showing three out of ten actions, (a) microwaving food, (b) picking objects, (c) unstacking objects, and example frames from three actions from the subset selected of the 51 actions in HMDB which include objects, (d) shoot gun, (e) draw sword, (f) kick ball.

Table 1: Comparison of different descriptors on the databases.

| Databases | CAD120 (%) | HMDB (%) |
|---|---|---|
| RGB trajectories | 32.30 | 38.13 |
| RGB HOG | 70.17 | 54.29 |
| RGB HOF | 49.02 | 41.6 |
| RGB MBH | 46.97 | 38.30 |
| RGB HOG3D | **83.94** | **71.98** |
| Depth trajectories | 56.34 | n/a |
| Depth HOG | 55.99 | n/a |
| Depth HOF | 56.47 | n/a |
| Depth MBH | 55.18 | n/a |
| FPFH | 60.51 | n/a |

Table 2: Context and modal influence on the databases using two approaches: ours (CMMKL) and uniformly weighted (UW) (Bautista-Ballester et al., 2014).

| CAD120 Database | UW (%) | CMMKL (%) | Kernel Weights |
|---|---|---|---|
| Object info. combined with RGB descriptor | | | |
| obj+RGB traj. | 36.34 | 56.57 | 0.5/0.7 |
| obj+RGB HOG | 75.21 | 86.32 | 0.2/0.8 |
| obj+RGB HOF | 54.78 | 74.94 | 0.4/0.8 |
| obj+RGB MBH | 54.50 | 68.32 | 0.4/0.6 |
| Depth info. combined with RGB descriptor | | | |
| Depth+RGB traj. | 72.61 | 83.19 | 0.7/0.3 |
| Depth+RGB HOG | 81.63 | 89.59 | 0.9/0.9 |
| Depth+RGB HOF | 75.08 | 86.53 | 0.4/0.8 |
| Depth+RGB MBH | 79.03 | 87.96 | 0.1/0.4 |
| 3D info. combined with RGB descriptor | | | |
| FPFH+RGB traj. | 62.03 | 87.98 | 0.6/0.4 |
| FPFH+RGB HOG | 69.98 | 90.67 | 0.2/0.5 |
| FPFH+RGB HOF | 67.65 | 89.28 | 0.7/0.7 |
| FPFH+RGB MBH | 69.27 | 90.18 | 0.8/0.7 |
| **HMDB Databases** | **UW (%)** | **CMMKL (%)** | **Kernel Weights** |
| Object info. combined with RGB descriptor | | | |
| obj+RGB traj. | 39.81 | 55.43 | 0.1/0.2 |
| obj+RGB HOG | 64.76 | 86.72 | 0.5/0.7 |
| obj+RGB HOF | 44.78 | 65.37 | 0.3/0.3 |
| obj+RGB MBH | 47.10 | 61.61 | 0.8/0.5 |

et al., 2011). Instead, we use the split in (Bautista-Ballester et al., 2014), which ensures the presence of as many variation as possible by following a proportion of clips similar to that in the complete database. These variations include what part of the body is shown, the number of people involved in the action, the camera motion and viewpoint, and the quality of the video.

The split consists of 6 different actions with 20 videos per action, resulting in 120 videos in total. These actions are *ride bike*, *shoot gun*, *shoot bow*, *draw sword*, *swing baseball* and *kick ball*. The purpose of this selection is dual: first, ensuring that an object always appears in the action, and second, ensuring the presence of as many variations as possible. Similar actions like *draw sword* and *swing baseball* are also taken into account, a fact that makes the set more challenging.

# 5 EVALUATION

In Section 5.1 we evaluate our CMMKL-SVM approach on CAD120 and HMDB datasets. We first evaluate single descriptors in order to find the most significative ones. Later, we evaluate the combination of different kernels and obtain the weights that informs us of the relevance of each kernel. In section 5.2 we compare our results for each database with those of the state of the art.

## 5.1 Evaluating CMMKL-SVM

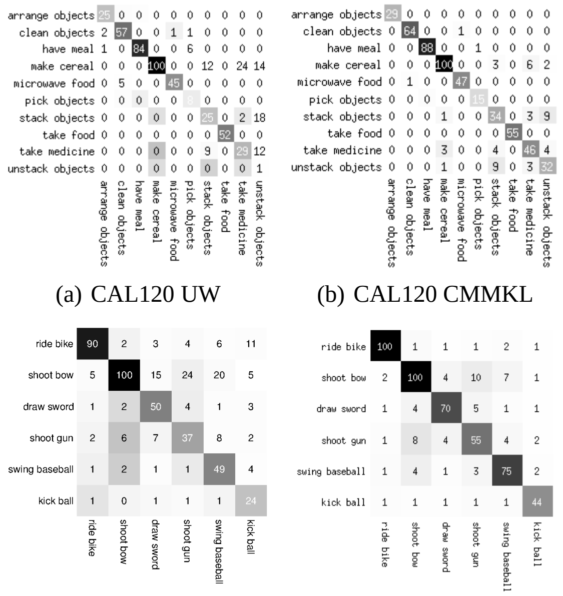
The use of CMMKL-SVM allows to add different descriptors into the standard BoW approach for action recognition. This approach permits the inclusion of several image descriptors into this scheme as explained in (Bautista-Ballester et al., 2014), and reduces the effect of information redundancy weighting a multikernel SVM. This approach improves the performance with respect to any singular descriptor or an averaged combination of them.

In our first experiment we calculate the average accuracy for each of the following descriptors: trajectories, HOG, HOF, MBH, HOG3D, Depth trajectories, Depth HOG, Depth HOF, Depth MBH, Depth HOG3D and FPFH. As we can see in Table 1, HOG3D descriptor gives the best action recognition performance. HOG3D avoid non-trivial preprocessing steps, such as tracking and segmentation, fuses 2D space and time information, and provides descriptors invariant to illumination and camera motion. This aspect shows that using a unique optimal descriptor can be better than a combination of several descriptors that perform worse individually. This is apparent in the fact that HOG3D obtains a 71.98% for HMDB and 83.94% for CAD120.

An extra objective of our approach is to overpass this performance by using CMMKL-SVM with the best weighted combination of descriptors using RGB videos, Depth videos, 3D points and objects. That would considerably reduce the time of the overall procedure, taking into account that HOG3D is quite computationally expensive. Additionally, Depth descriptors give similar results, (55%), for each single descriptor -trajectories, HOG, HOF and MBH- meaning that these descriptors lose their singular characteristics when used for depth videos. On the other hand, HOG and FPFH are the best choices when used as single descriptors, obtaining a recognition rate of 70.17% and 60.51% respectively in CAD120. This is due to the fact that they give spatial information of the action, a fact that has been verified in works like (Wang et al., 2013)(Bautista-Ballester et al., 2014).

In the second experiment, our purpose is to observe the influence of the context (objects) and mode (Depth, 3D) when employing single descriptors (trajectories, HOG, HOF, MBH) on RGB videos. The results are shown in Table 2. We perform the experiments with our approach CMMKL-SVM and the uniformly weighted approach in (Bautista-Ballester et al., 2014). We show that the fusion of context and mode information in a MKL framework is better than averaging kernels. *Having a look at the results in Table 2 we can observe that the addition of context, gives an important improvement of 20% on the average recognition rate for every trial in HMDB when using our approach*. For CAD120, this improvement is much lower than for HMDB, 10% on average, due to the quality of the videos and the lack of extensive variability in conditions such as illumination and viewpoint. In Table 2 we show how context, depth or 3D information always outperforms the recognition accuracy reached using a single descriptor.

The third experiment wants to find the best combination between all of the descriptors. The experiment has been performed choosing a first descriptor and progressively adding new ones in order to see the effect of the inclusion of this new information into the CMMKL-SVM. To see the best improvements, we have chosen the descriptor that contributes the least, i.e., trajectories. These results can be seen in Table 3. Any addition improve the results, but the question is which one provides the best results since adding new channels results in higher computational costs. Therefore, we want the least number of channels that provides the best results. We can conclude from these results that the addition of descriptors which provides redundant information leads to a lack of improvement. For example, the addition of HOF, object or FPFH to the combination trajectories + HOG leads to no significant improvement. We must observe that HOF provides temporal information in a similar sense as trajectories.



| | arrange objects | clean objects | have meal | make cereal | microwave food | pick objects | stack objects | take food | take medicine | unstack objects |
|---|---|---|---|---|---|---|---|---|---|---|
| arrange objects | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| clean objects | 2 | 57 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| have meal | 1 | 0 | 84 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| make cereal | 0 | 0 | 0 | 100 | 0 | 0 | 12 | 0 | 24 | 14 |
| microwave food | 0 | 5 | 0 | 0 | 45 | 0 | 0 | 0 | 0 | 0 |
| pick objects | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 |
| stack objects | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 2 | 18 |
| take food | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 52 | 0 | 0 |
| take medicine | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 29 | 12 |
| unstack objects | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

(a) CAL120 UW

| | arrange objects | clean objects | have meal | make cereal | microwave food | pick objects | stack objects | take food | take medicine | unstack objects |
|---|---|---|---|---|---|---|---|---|---|---|
| arrange objects | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| clean objects | 0 | 64 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| have meal | 0 | 0 | 88 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| make cereal | 0 | 0 | 0 | 100 | 0 | 0 | 3 | 0 | 6 | 2 |
| microwave food | 0 | 1 | 0 | 0 | 47 | 0 | 0 | 0 | 0 | 0 |
| pick objects | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 |
| stack objects | 0 | 0 | 0 | 1 | 0 | 0 | 34 | 0 | 3 | 9 |
| take food | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 55 | 0 | 0 |
| take medicine | 0 | 0 | 0 | 3 | 0 | 0 | 4 | 0 | 46 | 4 |
| unstack objects | 0 | 0 | 0 | 1 | 0 | 0 | 9 | 0 | 3 | 32 |

(b) CAL120 CMMKL

| | ride bike | shoot bow | draw sword | shoot gun | swing baseball | kick ball |
|---|---|---|---|---|---|---|
| ride bike | 90 | 2 | 3 | 4 | 6 | 11 |
| shoot bow | 5 | 100 | 15 | 24 | 20 | 5 |
| draw sword | 1 | 2 | 50 | 4 | 1 | 3 |
| shoot gun | 2 | 6 | 7 | 37 | 8 | 2 |
| swing baseball | 1 | 2 | 1 | 1 | 49 | 4 |
| kick ball | 1 | 0 | 1 | 1 | 1 | 24 |

(c) HMDB UW (Bautista-Ballester et al., 2014)

| | ride bike | shoot bow | draw sword | shoot gun | swing baseball | kick ball |
|---|---|---|---|---|---|---|
| ride bike | 100 | 1 | 1 | 1 | 2 | 1 |
| shoot bow | 2 | 100 | 4 | 10 | 7 | 1 |
| draw sword | 1 | 4 | 70 | 5 | 1 | 1 |
| shoot gun | 1 | 8 | 4 | 55 | 4 | 2 |
| swing baseball | 1 | 4 | 1 | 3 | 75 | 2 |
| kick ball | 1 | 1 | 1 | 1 | 1 | 44 |

(d) HMDB CMMKL

Figure 3: Confusion matrices for: (a) CAD120 database using objects, FPFH, Depth HOG, trajectories, HOG using UW approach (Bautista-Ballester et al., 2014) with average performance for 500 codewords: 79.73%, (d) CAD120 with our approach using the same configuration as (a), with average performance for 500 codewords: 90.83% (c) HMDB database using objects, trajectories, HOG, HOF, MBH descriptors as it is done in (Bautista-Ballester et al., 2014) with average performance for 500 codewords: 72.97%, (d) HMDB with our approach using the same configuration as (a), with average performance for 500 codewords: 85.41%.

## 5.2 Discussion

Comparing to the state-of-the-art, on one hand, (Koppula et al., 2013) obtained a 93,5% in CAD120 database using a CRF-based approach. We obtain a similar recognition accuracy of 92.83% using CMMKL-SVM. On the other hand, we significantly improve the results for HMDB, where (Bautista-Ballester et al., 2014) used a fusion of objects and RGB descriptors by averaging a multikernel SVM reaching 71,57%, much lower than our result of 85.41%. Table 4 shows this comparison for CAD120 and Table 5 for HMDB. The more realistic the database is the more relevant the acquisition and weights of contextual and multimodal information are.

Referring to Table 3, we can see the importance of weighting channels. Using a kernel averaging scheme (Bautista-Ballester et al., 2014) always obtained a lower performance than our approach, which takes into account the redundancy of information introduced by similar descriptors. This can be seen in

Table 3: Using different descriptors combinations on the databases with our approach.

| Database | CAD120 | | |
|---|---|---|---|
| | UW(%) | CMMKL(%) | Kernel Weights |
| trajectories + HOG | 71.04 | 83.24 | 0.2/0.6 |
| trajectories + HOF | 49.94 | 71.1 | 1.0/0.7 |
| trajectories + DepthHOG | 73.28 | 79.15 | 0.10/0.81 |
| trajectories + HOG + HOF | 75.10 | 85.94 | 0.5/0.5/0.1 |
| trajectories + HOG + obj | 71.40 | 83.60 | 0.8/0.9/0.4 |
| trajectories + HOG + FPFH | 71.90 | 90.33 | 0.8/0.4/0.6 |
| trajectories + HOG + HOF + MBH | 77.71 | 87.60 | 0.1/0.2/0.3/0.4 |
| trajectories + HOG + HOF + obj | 75.30 | 84.66 | 0.9/0.5/0.5/0.7 |
| trajectories + HOG + HOF + DepthHOG | **84.78** | 90.70 | 0.9/0.7/0.4/0.9 |
| trajectories + HOG + HOF + FPFH | 76.04 | 89.75 | 0.8/0.6/0.4/0.2 |
| trajectories + HOG + HOF + MBH + obj | 77.92 | 85.21 | 0.3/1.0/0.5/0.8/0.1 |
| trajectories + HOG + obj + FPFH + DepthHOG | 79.73 | **92.83** | 0.0/0.6/1.0/0.8/0.5 |
| Database | HMDB | | |
| | UW(%) | CMMKL(%) | Kernel Weights |
| trajectories + HOG | 57.83 | 82.24 | 0.3/0.4 |
| trajectories + HOF | 48.64 | 65.28 | 0.3/0.3 |
| trajectories + HOG + HOF | 64.67 | 85.82 | 0.1/0.8/0.3 |
| trajectories + HOG + obj | **71.57** | **85.84** | 0.3/0.7/0.1 |
| trajectories + HOG + HOF + MBH | 70.04 | 80.69 | 0.7/0.8/0.1/0.5 |
| trajectories + HOG + HOF + obj | 69.39 | 85.36 | 0.2/0.9/0.7/0.6 |
| trajectories + HOG + HOF + MBH + obj | 72.97 | 85.41 | 0.6/0.1/0.4/0.0/0.2 |

the combination trajectories + HOG + HOF, where trajectories almost loses its importance (0.1) because of other descriptors such as HOF (0.3), which provides temporal information like trajectories. However, HOG still remains the most significant descriptor (0.8). This reinforces the hypothesis made in (Bautista-Ballester et al., 2014) that the strongest descriptors are those that provide spatial information.

Finally, regarding the confusion of the actions, CMMKL-SVM reduces confusion between actions, even for similar actions, as can be seen in Fig. 3. For example, *Unstacking objects* for CAD120 is easily confused with *Stacking objects*, a relation that averaging kernels cannot break (1%) but our approach does (32%). The same happens when *kicking a ball*, where averaging kernels performs a 24% and CMMKL a 44%. In general, all actions in both databases have their confusion index reduced. Therefore, the overall performance of our action recognition approach is higher than other state-of-the-art approaches.

Table 4: Comparison to the state of the art on CAD120 database.

| Work | Approach | Avg. acc. |
|---|---|---|
| (Koppula et al., 2013) | CRF-based | 93.50% |
| Ours | CMMKL-SVM | **92.83**% |

## 6 CONCLUSIONS

In this paper we have proposed a methodology to combine different descriptors within a standard action recognition scheme based on BoW. Our approach adds information related to the objects, depth maps and 3D points, and shows an increment of the overall action recognition performance. The addition of the extra image descriptors, either from RGB context or sensor modality, leads to an increment of the computational cost. As a consequence, it is important to discriminate, or even discard, the less important descriptors. Our approach complements space and time information extracted with video descriptors, and proposes a procedure to incorporate and weight any contextual and modal information that can be further generalized to include other data provided by new context descriptors and/or new devices. Additionally, the present approach also shows that the best results are

Table 5: Comparison to the state of the art on HMDB database.

| Work | Approach | Avg. acc. |
|---|---|---|
| (Bautista-Ballester et al., 2014) | Multichannel UW | 71.57 % |
| Ours | CMMKL-SVM | **85.84**% |

obtained when kernels from spatial, temporal, context, 3D points and depth are combined within the CMMKL-SVM approach. In this respect, the highest recognition rates (92.83%) have been obtained when a combination of trajectories, HOG, FPFH, Depth and object is used. Due to the relevant importance to intelligent robots, our future work will focus on the improvement of multimodal fusion and the reduction of the computational burden by exploiting different optimization techniques for MKL, allowing a quicker response of the robot to interact with humans by either imitating or anticipating actions.

## ACKNOWLEDGEMENTS

This research has been partially supported by the Industrial Doctorate program of the Government of Catalonia, and by the European Community through the FP7 framework program by funding the Vinbot project (N 605630) conducted by Ateknea Solutions Catalonia.

## REFERENCES

Bautista-Ballester, J., Vergés-Llahí, J., and Puig, D. (2014). Using action objects contextual information for a multichannel svm in an action recognition approach based on bag of visual words. In *International Conference on Computer Vision Theory and Applications*, VISAPP.

Bilinski, P. and Corvee, E. (2013). Relative Dense Tracklets for Human Action Recognition. *10th IEEE International Conference on Automatic Face and Gesture Recognition*.

Bucak, S., Jin, R., and Jain, A. (2014). Multiple kernel learning for visual object recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1354–1369.

Bucak, S., Jin, R., and Jain, A. K. (2010). Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition. In *Advances in Neural Information Processing Systems*, pages 325–333.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Gehler, P. and Nowozin, S. (2009). Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 2836–2843. IEEE.

Ikizler-Cinbis, N. and Sclaroff, S. (2010). Object, scene and actions: Combining multiple features for human action recognition. In *Proceedings of the 11th European Conference on Computer Vision: Part I*, ECCV'10, pages 494–507, Berlin, Heidelberg. Springer-Verlag.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004.

Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72.

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

Pieropan, A., Salvi, G., Pauwels, K., and Kjellstrom, H. (2014). Audio-visual classification and detection of human manipulation actions. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 3045–3052.

Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*.

Rusu, R. B. (2009). *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany.

Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA. ACM.

Tsai, J.-S., Hsu, Y.-P., Liu, C., and Fu, L.-C. (2013). An efficient part-based approach to action recognition from rgb-d video with bow-pyramid representation. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2234–2239.

Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A. (2009). Multiple kernels for object detection. *In Proceedings of the International Conference on Computer Vision, 2009.*

Wang, H., Kläser, A., Schmid, C., and Liu, C. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2011). Action Recognition by Dense Trajectories. In *IEEE Conf. on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States.

Wang, H. and Schmid, C. (2013). Action Recognition with Improved Trajectories. In *ICCV 2013 - IEEE International Conference on Computer Vision*, pages 3551–3558, Sydney, Australie. IEEE.