**Machine Vision & Applications manuscript No.**
(will be inserted by the editor)

# Human Interaction Categorization by Using Audio-Visual Cues

**M.J. Marín-Jiménez · R. Muñoz-Salinas · E. Yeguas-Bolivar · N. Pérez de la Blanca**

Received: date / Accepted: date

**Abstract** Human Interaction Recognition in uncontrolled TV video material is a very challenging problem because of the huge intra-class variability of the classes (due to large differences in the way actions are performed, lighting conditions and camera viewpoints, amongst others) as well as the existing small inter-class varibility (e.g. the visual difference between hug and kiss is very subtle). Most of previous works have been focused only on visual information (i.e. image signal), thus missing an important source of information present in human interactions: the audio. So far, such approaches have not shown to be discriminative enough.

This work proposes the use of *Audio-Visual Bag of Words* (AVBOW) as a more powerful mechanism to approach the HIR problem than the traditional *Visual Bag of Words* (VBOW). We show in this paper that the combined use of video and audio information yields to better classification results than video alone. Our approach has been validated in the challenging TVHID dataset showing that the proposed AVBOW provides statistically significant improvements over the VBOW employed in the related literature.

**Keywords** Human Interactions · Audio · Video · BOW

M.J. Marín-Jiménez · R. Muñoz-Salinas · E. Yeguas-Bolivar
Department of Computing and Numerical Analysis.
University of Córdoba.
14071 Córdoba (Spain)
E-mail: {mjmarin,rmsalinas,eyeguas}@uco.es

N. Pérez de la Blanca
Department of Computer Science and Artificial Intelligence
University of Granada.
18071 Granada (Spain)
E-mail: nicolas@ugr.es

## 1 Introduction

Given a video clip where there are people interacting between them, the goal of this work is to automatically assign a single category label – from a set of predefined ones – to such human interaction. We address this problem by considering human interactions as an *audio-visual event*, i.e. sequence of image frames plus sound (see Fig. 1).

In Fig. 2 we can see four scenes of people interacting. In such scenes, there are two people very close with the arms holding the other person. Two of such scenes – extracted from TV Human Interactions Dataset (TVHID) [20] – have the label *kiss* and the other ones the label *hug*. In spite of the fact that, nowadays, recorded video clips contain not only image but also sound, the current approaches for distinguishing such kind of human interactions only make use of the video pixels, discarding the rich information encoded in the audio signal. The previously presented cases can be clearly ambiguous for a computer if we only take into account the visual information. However, if we focus on the audio signals represented in Fig. 3, we notice that *kiss* and *hug* have different audio patterns. Furthermore, many human interactions have associated very well defined audio-visual patterns – words as *hi, hello, nice* or *meet* are very common during a *hand-shake* – introducing a very clear discrimination with other interactions. Therefore, in this paper we introduce a new approach to deal with the categorization of human interactions by using audio-visual information.

Our contribution is two-fold: (i) we introduce the use of the audio signal in the challenging problem of human interaction categorization; and, (ii) we carry out a thorough experimental study on TVHID where it is shown that the combination of visual and audio infor-

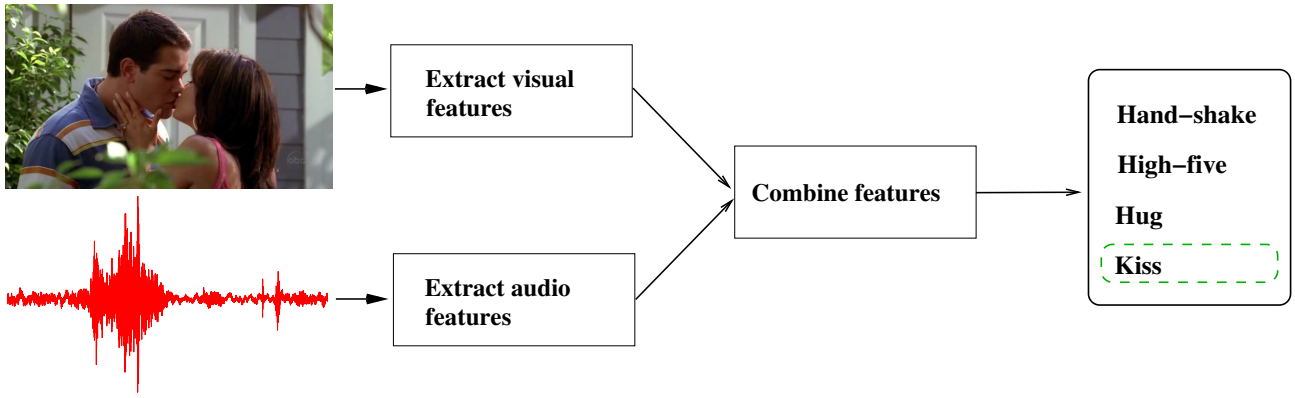


Fig. 1: **Proposed pipeline for human interaction categorization in TV shows**. Audio and visual information combined in an unified framework in order to distinguish a human interaction.

mation offers better results than only using the visual one – as done up to this moment.

The rest of the paper is organized as follows. Sec. 1.1 explains some of the more relevant works related to ours. In Sec. 2 we introduce the audio-visual model used in our proposal, which is based on the successful Bag of Words model. The experiments and results are presented in Sec. 3. The paper is concluded in Sec. 4.

### 1.1 Related Works

In recent years an increasing number of research papers have been published in the context of Human Action Recogniton (HAR) in videos. For example, [30] compiled published works within a period of 20 years devoted to human actions and activities. In the early years the proposed approaches were tested on *artificially generated* datasets [7,25,37], where a single person performed a target action (e.g. walk, jump, hand-wave,...) in controlled scenarios. Soon, *realistic* datasets were compiled from Hollywood movies [11,12,18], where one or more persons perform a named action in an uncontrolled, and usually cluttered, scenario. A particular case of human actions is *human interactions*. We can distiguish between the interactions performed by a person with an object – as *smoking a cigarette* [12] or *playing a musical instrument* [6,39] – or between two or more persons, as *hand-shaking* or *hugging* [20,24].

Human Interaction Recognition (HIR) in video sequences [17,20,23,24] is a very difficult problem due to several reasons: (a) action performance and camera viewpoint – the different velocities and manners of performing the interaction by the persons in combination with diverse camera viewpoints; (b) imaging conditions – the ever-present difficulties found when working with images from real scenarios (i.e. uncontrolled imaging conditions); (c) non-stationary noise – cluttered and different backgrounds, partial occlusions or diverse person clothing; and, (d) relative volume occupied by the interaction – only a very small region of the pixels along with a short number of video frames are related to the event of interest (e.g. the involved hands in *hand-shaking*). The latter reason is the one that mostly differentiates realistic human interactions in video with regard to still images or simulated human actions (e.g. *jumping* in Weizmann dataset [7] is very repetitive). In comparison with the task of object/concept categorization on still images, where the area of interest is a large percentage of the image, the HIR problem is clearly much more challenging. Also if we compare with the HAR problem, we see HIR more challenging not only due to higher complexity but also due to the difficulty of getting large training databases from real scenes. Up to our knowledge, the only existing dataset devoted to human interactions in realistic situations is TV Human Interactions Dataset (TVHID), introduced by Patron-Perez *et al.* [20]. In [20], the problem of HIR on this dataset is addressed by firstly detecting and tracking people and, then, by combining head pose estimators with visual local context descriptors (i.e. Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) features).

There are some papers where the problem of semantic video retrieval is addressed by using only audio features. For example, Bakker and Lew [1] combine local and global audio features to classify sound samples from video into several classes as, for example, *speech*, *music*, *automobile* or *explosion*. Tzanetakis and Chen [31] build audio classifiers to distinguish between *male voice*, *female voice*, *noise*, *music* and *silence* from videos. Bredin *et al.* [2] approach the problem of content-based video


Fig. 2: **Kiss or hug?** Sometimes visual information on its own is not enough to automatically distinguish between human interactions. In this figure, (a) and (c) correspond to *hug*, whereas (b) and (d) correspond to *kiss*. (See Fig.3 for a graphical representation of their associated audio signal.)

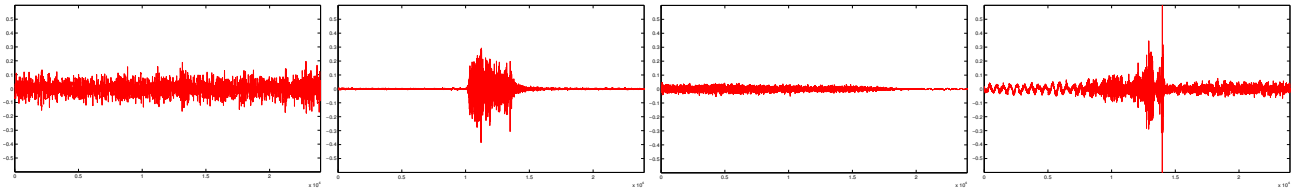

Fig. 3: **Audio signals for scenes in Fig.2.** Note the representative peaks in the audio signal for the *kiss* interaction examples. Such kind of peaks are not present, for example, in a *hug* interaction.

retrieval by combining multiple audio classifiers in a HMM-based framework.

In McCowan *et al.* [19] it is shown how the use of audio-visual events can improve the recognition of group actions in meetings within controlled scenarios. However, we approach the uncontrolled case in this paper. On the other hand, in recent years *concepts* (e.g. news, commercials, sports,...) are assigned to videos by using the combination of visual, audio and even textual information [28]. For example, in [21], [8] and [14], image (e.g. SIFT, HOG, Gist) and audio (e.g. MFCC, WASF) descriptors are combined by using different approaches for the task of multimedia event detection. Amitha *et al.* [21] propose and evaluate two types of fusion: (a) training high-level classifiers on the output of previously trained feature-specific classifiers, and, (b) learning a linear combination of low-level classifiers. In order to represent multimedia events, Inoue *et al.* [8] use Gaussian Mixture Models and Support Vector Machines (SVM) to combine audio and visual features. Sidiropoulos *et al.* [26] introduce the usage of audio in the problem of video scene segmentation. Recently, in Jiang *et al.* [9] a new challenge for multimedia video classification is proposed. However the focus is not on human interaction but on event classes. In addition, and as mentioned above, in our case only a short and small part of the signal helps to classify the whole video sequence. Despite these related works, audio has not been exploited on HIR yet, what is one of the main novelties of this paper.

## 2 Representation model

Inspired in the models used by the *document retrieval* community, Sivic and Zisserman [27] proposed an analogy between the textual words and the visual words (i.e. image region descriptors) with the idea of representing an image (i.e. the *document*) as an orderless collection of visual words: a *Bag of Words* (BOW).

In its simplest way, a BOW is equivalent to a histogram $\mathbf{h}$ with $K$ bins (i.e. as much as words in the dictionary $\mathbf{D}$) where each bin represents how many times a visual word is present in the target image. In general, the histogram is L1 normalized.

The operation of assigning a word to a bin histogram, implies the process of finding the word $\mathbf{D}^{(j)}$ that makes minimum the distance between the current word and all the words included in the dictionary. Euclidean distance is a common choice to carry out the word assignment.

Although this representation was originally used on images, it was generalized in the recent years to describe video sequences [11]. Fig. 4 shows the classical pipeline used to learn representations of human actions: (i) compute Spatio-Temporal Interest Points (STIP) on input video; (ii) compute descriptors from STIP (e.g. HOG/HOF); (iii) learn a dictionary of visual words from the set of STIP extracted from the training videos; (iv) describe the videos by using the STIP descriptors and the previously learnt dictionary; and, (v) train a discriminative classifier (e.g. SVM).

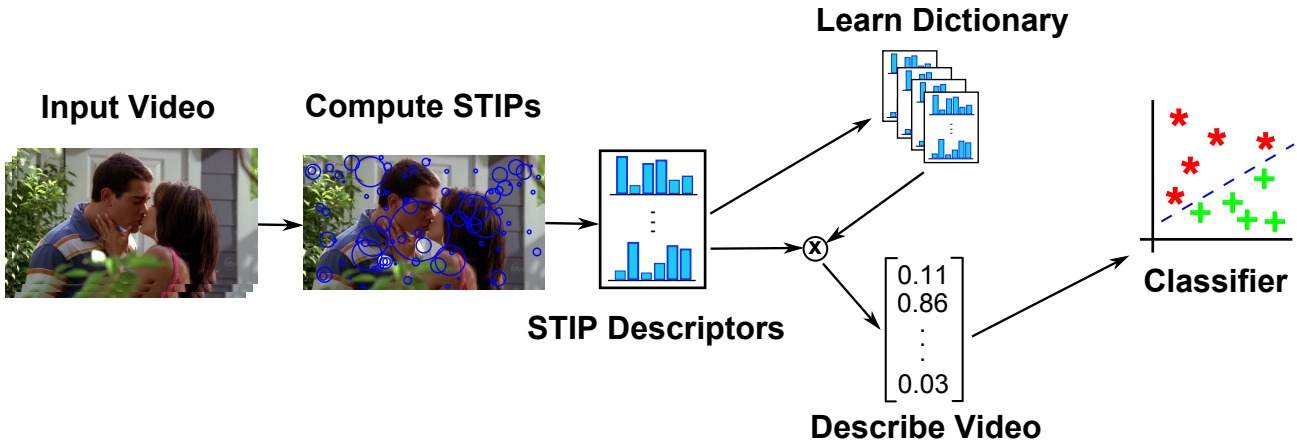


Fig. 4: **Classical pipeline for learning with BOW**: (i) Compute STIP on input video; (ii) compute descriptors from STIP; (iii) learn a dictionary of visual words; (iv) describe video by using STIP descriptors and learnt dictionary; and, (v) train a discriminative classifier.

For a given video sequence, we build different BOWs depending on the kind of *words* used: visual or audio descriptors.

We use the popular $K$-means algorithm [15] to build a dictionary $\mathbf{D}$. The goal of $K$-means clustering is to find a partition of the descriptor space in $K$ regions. Each region will be represented by the mean vector of its components. We have chosen the implementation of this algorithm included in VLFeat library [33].

The resulting audio-visual video descriptor will be used as input for a classifier. In our case, we have chosen a Support Vector Machine (SVM) with $\chi^2$ kernel, which has shown to be very effective when working with histogram-based representations [34].

### 2.1 Visual features

Spatio-Temporal Interest Points (STIP) were firstly introduced by [10] and applied to the problem of recognizing individual human actions (e.g. walk) in video. They propose a *Harris3D* operator to detect *salient* points in the space-time volume. In addition to the $(x, y, t)$ coordinates, each STIP has associated a spatial and a temporal scale $(\sigma_s, \sigma_t)$ that delimit the video volume where the event of interest happens. An effective alternative to Harris3D operator is a simple *dense sampling*. It consists of extracting video blocks at regular locations and scales in space and time, usually, with overlapping. In several problems, this approach has shown state-of-the-art results [36].

The most popular volume descriptors [11] for STIP are Histograms of Oriented Gradients [3] (HOG) and Histograms of Optical Flow [4] (HOF). HOG encodes local appearance whereas HOF encodes local motion.

In order to compute a HOG descriptor, the image volume is divided into a a $n_x \times n_y \times n_t$ dense grid of *cells*, where each cell will contain a local histogram over orientation bins. Then, at each pixel, the image gradient vector is computed. Each pixel votes into the corresponding orientation bin with a vote weighted by the gradient magnitude. The votes are accumulated over the pixels of each cell. Afterwards, in order to provide illumination invariance, a normalization stage is performed over each *block* (group of cells). The normalized histograms of all of the blocks are concatenated to build the final HOG descriptor. A similar procedure is used for computing HOF descriptor, but replacing image gradient (spatial) by optical flow.

For our experiments, 4 orientation bins will be used for HOG and 5 for HOF.

### 2.2 Audio features

In order to use the audio signal in the BOW framework, firstly, we split the audio signal into overlapping *frames* of $t$ seconds (i.e. $t = 0.05$). An example over an audio signal extracted from a *kiss* example is represented in Fig. 5. Then, we compute on each audio frame a set of descriptors.

The simplest descriptor is the *raw* signal per se (i.e. the actual values), which will be used in the experimental section as baseline on audio features.

The use of Mel-frequency cepstral coefficients (MFCC) as audio descriptor is a popular choice specially in the fields of speech or music recognition [5,13]. It offers a

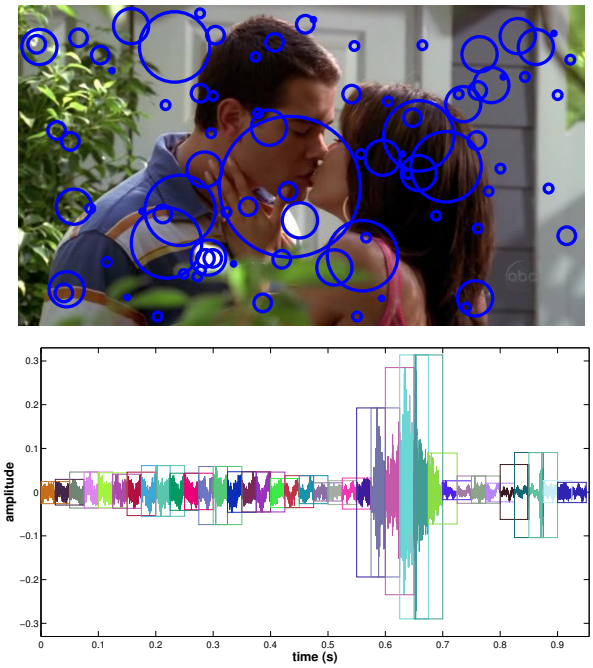


Fig. 5: **Local audio-visual features. (top)** Spatio-Temporal Interest Points (STIP) are used as basis for visual features. A HOG/HOF descriptor is computed for each STIP. **(bottom)** Audio signal is divided into overlapping *frames*. In this example, the signal is divided in frames of 0.05 seconds overlapping 0.025 s. Features are extracted over each audio frame. Then, each resulting feature vector is assigned to a *word* of the corresponding dictionary.

description of the spectral shape of the audio in a given interval of time. It is computed as follows [13]:

1. compute the Fourier Transform (FT) of the signal;
2. map the powers of the spectrum obtained with FT onto the *mel scale* (i.e. perceptual scale of pitches [29]);
3. compute the logs of the powers at each of the mel frequencies;
4. compute the Discrete Cosine Transform of the list of mel log powers, as if it were a signal.

The amplitudes of the resulting spectrum define the MFCC.

In addition to MFCC, and for comparison purposes, we extract the following set of simple features in the time domain [13]:

- zero-cross: the number of times the signal changes sign (i.e. crosses $X$-axis);
- coefficient of skewness:

$$\mu_3/\sigma^3$$

, where $\mu_3$ is the third-order moment of the data and $\sigma$ is its standard deviation;
- excess kurtosis:

$$\mu_4/\sigma^4 - 3$$

, where $\mu_4$ is the fourth-order moment of the data and $\sigma$ is its standard deviation;
- flatness: the flatness of the data results from the ratio between the geometric mean and the arithmetic mean;
- entropy: the relative Shannon's entropy of the data (i.e. it is divided by the length of the signal).

For each audio frame, the previous features are concatenated into a single feature vector which will describe such audio frame.

### 2.3 Audio-Visual Bag of Words

The combination of audio and video information has been previously employed in video categorization [9], however, its use in the problem of interaction recognition has not been yet deeply explored. The action recognition problem normally involves a single person, and people do not usually speak to themselves while performing actions. On the other hand, interaction involves two or more people, and both visual and audio information plays important role to communication. This work aims at showing that the combination of both sources of information (see Fig. 1) can yield to better results in this problem, than their standalone use.

Two main approaches in data fusion can be considered, *early fusion* and *late fusion*. In the first approach, fusion is performed before the classification process takes place. Normally, it consists in joining all the features into a single feature vector. Late fusion, on the other hand, performs first classification of all sources of information separately, and then, fuses the results. Most often, another classifier is trained on the output of the individual classifiers. This work tests both approaches in order to analyze their performance.

## 3 Experiments and results

This section explains the experiments performed to validate our proposal. Our goal is to demonstrate that audio information can be employed to improve the classification performance in the HIR problem. To do so, we have first tested the performance of video features. In our work, HOG and HOF features have been tested both separately and together. Statistical tests have been run on the results so as to analyze which combination performs better. Then, we have tested performance

Fig. 6: **TV Human Interactions Dataset: *hand-shake*, *high-five*, *hug* and *kiss***. The different viewpoints and challenging imaging conditions (lighting, cluttered background, clothing, partial occlusions...) make their recognition with only visual information a very difficult problem.

of the audio features previously explained. Finally, we have tested the combined use of audio and visual features. Again, statistical tests have been run to analyze the impact of the combination. With regard to the feature combination method, early and late fusion approaches are evaluated, with special emphasis in early fusion.

Experimentation has been carried our in the TV Human Interactions Dataset (TVHID) [20] which consists of 200 videos from TV shows grouped in 4 categories: 50 *hand-shake*, 50 *high-five*, 50 *hug* and 50 *kiss*. In addition, a set of 100 *negative* videos (i.e. none of the other interaction categories) is included. Fig. 6 contains examples of the four interactions included in TVHID. Note the different imaging conditions (e.g. illumination, scale, background clutter,...) where the interactions happen. Each video clip is labelled with a single interaction class from the possible ones. The dataset provides information about the frame intervals where the interaction happens within each video plus additional information such as the coordinates of upper-body bounding boxes and an approximation of the head orientations.

The rest of this section is structured as follows. Firstly, Sec. 3.1 explains the evaluation protocol and experimental setup employed. Then, we test the performance of visual (Sec. 3.2) and audio (Sec. 3.3) features independently. Finally, Sec. 3.4 shows the results of combining both sources of information.

### 3.1 Evaluation protocol and Experimental setup

Our proposal is evaluated in the context of *human interaction categorization*, i.e., given an input video, it must be classified into the correct category. Thus, it is a multiclass problem that has been addressed by training 4 *one-vs-all* binary classifiers. SVM with approximated $\chi^2$ kernel [34] are used in all our experiments, but in the ones of Sec. 3.4.2.

The TVHID data set is divided in two standard partitions that have been respected to allow a direct comparison with future and past results. So, training is first performed on one partition and test on the other one, and then the process is repeated by interchanging the role of the partitions. As measurement of performance $Succ$, we report the averaged percentage of correctly categorized test videos on the two trials (i.e. 2-fold cross-validation):

$$Succ = 100 \cdot \left( \frac{c_1}{n_1} + \frac{c_2}{n_2} \right) \quad (1)$$

where $c_1$ and $c_2$ are the number of correctly categorized videos in the first and second partitions, respectively, and $n_1$ and $n_2$ are the total number of evaluated videos during test time on each partition, respectively.

With regard to the image signal, we extract STIP only from the frame intervals where the interaction happens, discarding the STIP whose center is outside the *person region*. Such region is defined by computing the minimum and maximum $x$ from the upper-body bounding box coordinates of the annotated persons in the target frame. All the frame height is included in the person region. Since in this work we are mostly interested in the contribution of the audio features to HIR problem, we have adopted this preprocessing stage during training in order to minimize the noise that could be introduced in the evaluation by the visual regions located outside the *person region*. On the other hand, the audio signal (used both for training and testing) is extracted from the time interval where the interaction happens, as indicated by the dataset annotations.

In order to analyze the performance of the different features, statistical hypothesis tests [22] have been employed. Comparing exclusively the best results obtained by two set of features does not provide enough support to say whether the differences are significant.

Statistical hypothesis tests, in general, answer the question: *Assuming that the null hypothesis, $H_0$, is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?* That probability is known as the $p$-value and the null hypothesis is to consider that the features performances are equal. If the test proves that the null hypothesis is false, then, the differences observed are not due to chance but statistically significant. The reduced number of samples employed in our tests makes it difficult to determine their distribution. Therefore, non-parametric tests have been employed, since they do not require the assumption of normality or homogeneity of variance. Their main disadvantage (compared to parametric tests) is that for the same number of observations, they are less likely to lead to the rejection of a false null hypothesis. The hypothesis verified by all tests are $H_1$: the median difference can be considered statistically significant (not by chance); and $H_0$: otherwise. In all our tests, we have assumed $p = 0.05$.

Two different tests have been employed depending on the type of data, namely, the Mann-Whitney [16] and Wilcoxon signed-rank [38] test. The former will be employed for assessing whether two samples of independent observations tend to have larger values than the other. The latter, a paired test, analyzes the impact of an experiment on a population by measuring features before and after the experiment. In our case, the paired test will tell us if adding a feature to another has any impact on the classification results, e.g. adding audio to the video features.

### 3.2 Baseline: visual features

In this first experiment, we establish the baseline results obtained with BOW combined with STIP-based features (see Sec. 2.1). Different values of dictionary size $K$, in the range $[100, 2000]$ are tested (values of $K$ out of this range did not show any improvement), in addition to the use of Harris3D interest point detector and dense sampling. A maximum of $10^5$ randomly selected descriptors are used as input for the dictionary learning stage.

We have tested both HOG and HOF descriptors separately, and joined so as to analyze the impact of its combination. The HOG descriptor is a vector of 72 dimensions, whereas HOF descriptor has 90 dimensions. Therefore, the combined HOG+HOF descriptor has 162 dimensions.

Tab. 1 contains a summary of the results of this experiment. The *Succ* value for each configuration is reported. Keyword *dense* indicates that STIP have been extracted by using dense sampling, otherwise, Harris3D detector has been used.

Table 1: **Human interaction categorization on TVHID by using visual information.** Percentage *Succ* of correct categorization. The best performance for each descriptor is marked in bold.

| K/ features | *HOG* | *HOF* | *HOG+HOF* |
|---|---|---|---|
| 100 | 39.5 | 38.5 | 42.0 |
| 500 | 33.0 | **45.0** | **46.0** |
| 1000 | 36.5 | 43.0 | 42.5 |
| 2000 | 36.5 | 40.0 | 44.0 |
| 1000+dense | **39.5** | 43.0 | 44.5 |
| 2000+dense | 39.5 | 39.5 | 45.5 |

Table 2 shows the results of the tests carried out on the database with only visual features. For the comparison HOF vs HOG, we have employed the Mann-Whitney Test [16], while for the tests HOG vs HOG+HOF and HOF vs HOG+HOF we have employed the Wilcoxon signed-rank test [38].

Table 2: **Statistical analysis of the performance of visual features.** Values in brackets $(\mu, t)$ represent the average difference between the sets and the valid hypothesis.

| | HOF | HOG+HOF |
|---|---|---|
| HOG | $(+4.08, H_1)$ | $(+6.58, H_1)$ |
| HOF | – | $(+2.50, H_0)$ |

Each cell of the table shows the statistical comparison between their intersecting features. The values into the brackets are the average difference between the sets; and the hypothesis verified by the tests ($H_1$ if the median difference can be considered statistically significant, and $H_0$ otherwise). The tests have been conducted considering the column feature as first set, and the row feature as the second set. So, positive values for the average indicates that the column feature performs better than the row feature (e.g. in Tab. 2, HOG+HOF performs better than HOG). In all our tests, we have assumed $p = 0.05$.

The tests indicate that, using HOG, the mean success is 4.08 higher than using HOF, and that the improvement observed is not due to chance, but statistically significant, i.e., $H_0$ has only a probability of $p = 0.05$ of being true. It can also be observed that HOG+HOF obtains statistically significant differences when compared to HOG alone. With respect to HOG+HOF vs HOF, we observe an increase in the success, but we do not have enough support to indicate that their differences are statistically significant given the observations. From the results obtained, we conclude that the combination $HOG + HOF$ is the best video feature.

### 3.3 Evaluation of audio features

In this experiment we evaluate the use of audio features. For this experiment we employ the audio features introduced in Sec. 2.2: group $A1$ is composed by zero-cross, excess kurtosis, coefficient of skewness, flatness and entropy; $A2$ corresponds to mel spectrum (i.e. MFCC before DCT); and, $A3$ corresponds to MFCC. The feature vector $A1$ has 5 dimensions whereas vector $A2$ has 40 dimensions and $A3$ has 13 dimensions. As baseline feature, we have chosen the *raw* audio signal.

A maximum of $10^5$ randomly selected descriptors are used as input for the dictionary learning stage. We have tested different values of dictionary $K$ in the range $[25, 500]$. Values of $K$ out of that range did not show any improvement over the results reported in this paper.

Table 3: **Evaluation of audio features on the TVHID positive classes.** Percentage of correct categorization for the most representative configurations. $Ax$ indicates the group of audio features used in the experiment. *raw* refers to the actual audio signal. The best overall performance is marked in bold.

| K/ features | *raw* | $A1$ | $A2$ | $A3$ |
|---|---|---|---|---|
| 25 | 30.0 | 37.0 | 37.5 | 41.0 |
| 50 | 39.0 | 40.5 | 38.0 | **48.5** |
| 100 | 32.0 | 41.0 | 41.5 | 47.0 |
| 200 | 41.0 | 31.0 | 41.5 | 38.0 |
| 300 | 32.0 | 34.0 | 39.0 | 39.5 |
| 400 | 32.5 | 40.5 | 40.5 | 42.0 |
| 500 | 36.0 | 41.5 | 38.0 | 44.0 |

The results of the experiments are summarized in Tab. 3, that shows the *Succ* for each configuration. Using the results reported in the previous table, we have conducted the Mann-Whitney Test tests to compare the performance of the different audio features (see Tab. 4). As can be seen, the tests show that $A2$ and $A3$ present statistically significant differences with respect to the raw data. However, with the tests performed, it cannot be stated that there are significant differences between the $A1$, $A2$ and $A3$ features. Nonetheless, the best average results are obtained by the $A3$ set.

Table 4: **Statistical analysis of the performance of audio features**. Values in brackets $(\mu, H)$ represent the average difference between the sets and the valid hypothesis.

| | A1 | A2 | A3 |
|---|---|---|---|
| raw | $(+2.66, H_0)$ | $(+4.33, H_1)$ | $(+7.75, H_1)$ |
| A1 | – | $(+1.66, H_0)$ | $(+5.08, H_0)$ |
| A2 | – | – | $(+3.41, H_0)$ |

### 3.4 Feature combination

In this section we evaluate different ways of combining audio-visual features. In many problems the feature combination from several modalities improves the results from the best single modality. Here we conduct experiments to confirm this fact in our task at the same time that identifying the possible causes of the improvement. To do this we run experiments from a baseline early fusion technique using the simple concatenation of the audio and video features. Then, we compare the results with the state of the art technique for modality fusion, Multiple Kernel Learning (MKL) [35,32], and also with a technique based on a bi-modal codebook [40].

#### *3.4.1 Fusion baseline*

In this experiment, we compare the proposed audio-visual framework with the classical visual approach (see Sec. 3.2) employing an early fusion. Additionally, we aim at quantifing the impact of adding audio information to video in the sequences tested.

To that end, we have performed a thorough analysis of our proposal by combining the different HOG+HOF visual features ($\{K100, K200, \ldots, K2000{-}dense\}$) with all the audio combinations evaluated in Sec. 3.2 (i.e. $\{K25, K50, K100, \ldots, K500\}$) using three different early fusion approaches.

As baseline fusion method we choose the concatenation of the feature vectors. This model is equivalent to consider a linear combination of kernels with equal weights. SVM are trained on the concatenated feature vectors.

The results of this experiment are summarized in the rows $BLF$ of Tab. 5. Columns labeled as $Kx$ represent the HOG+HOF visual features, while rows represent audio features and a particular early fusion method. In each row we present the results of the Wilcoxon paired test that compares the results of the column visual feature vs audio-visual features.

For instance, the cell ($BLF$-$A1$,$K100$) shows the results of the Wilcoxon test when the HOG+HOF classifier with 100 features is compared to the all the classifiers that results from adding the audio features $\{A1-25,\ldots,A1-500\}$ (in total 7 different audio-visual classifiers). As a consequence, since $H_1$ holds true in this case, it means that the addition of the audio feature $A1$ (in general) obtains better results than video $K100$ alone.

The three pieces of data $(d, H, m)$ in the cells represents the following. First, $d$ is the average mean difference between the classifiers, where positive values indicate that audio-visual classifiers are better than the visual ones. Second, the $H$ denotes the most likely hypothesis (i.e. $H_1$ indicates that the difference is really significant). Finally, $m$ represents the average on the performance $Succ$ of the audio-visual classifiers.

### *3.4.2 Multiple Kernel Learning*

In this experiment we evaluate a Multiple Kernel Learning (MKL) [32] approach for early feature fusion. Let $((\phi_k(\mathbf{x}_i), y_i), i = 1,\cdots,N)$ be a sample from each one of the $K$ input feature descriptors $\phi_k$, where $y_i$ represents the class label. Let $f_1,\ldots,f_K$ be $K$ associated distance functions, where $f_k = \mathbf{w}_k^T\phi_k$. Then, the goal of the linear MKL is to find the optimal descriptor's kernel $\mathbf{K_{opt}} = \sum_k d_k\mathbf{K}_k$ where $\mathbf{K}_k$ is the $k$-th kernel matrix (i.e. function of $f_k$) and $\mathbf{d}$ are the weights. The estimation is carried out in as an SVM optimization framework where the primal problem can be formulated as:

$$\underset{\mathbf{w}_k,b,\mathbf{d},\xi\geq 0}{\text{Min}} \quad \frac{1}{2}\sum_k \frac{\mathbf{w}_k^t\mathbf{w}_k}{d_k} + C\sum_k \xi_k + \frac{\lambda}{2}||\mathbf{d}||_p^2 \qquad (2)$$

$$s.t.\ \ y_i\left(\sum_k \mathbf{w}_k^T\phi_k(\mathbf{x}) + b\right) \geq 1-\xi_i,\ \ i = 1,\cdots,N$$

where $||\cdot||_p$ represents the Euclidean $p$-norm. Nevertheless, this formulation is too simple for some applications since it is equivalent to concatenate the $K$ descriptors of each sample. A richer representation is proposed in [32] using the product of kernels instead of the sum. We have used both of it in our experiments. A $\chi^2$ distance and a product of exponential kernels of precomputed distance matrices with SVM classifiers have been used as feature distance and generalized kernel respectively. The results achieved with this early fusion method are summarized in the rows $MKL$ of Tab. 5.

### *3.4.3 Bi-modal codewords*

In contrast to the MKL approach, where a sophisticated combination method is used to fuse the information from each modality, in [40] a new way of fusing audio and video features is proposed creating audio-visual patterns represented in a bi-modal codebook. In short, this technique starts creating a bag of words model from the audio and the video modalities and then a distance matrix between the codewords of both dictionaries is estimated. In order to estimate the subset of codewords that explains the best the audio and video correlation, a spectral clustering technique is applied. The new subsets of features given by the clusters are used to define a bi-modal dictionary used to code the original audio and video codebooks. The *average*, *max* and *hybrid* criteria suggested in [40] to make the final coding have been tested.

In our case, the *max* criterium showed the best results with a bi-modal dictionary size of 50% of the original size. The results are summarized in rows *Bimodal* of Tab. 5.

### *3.4.4 Late fusion*

As commented in Sec. 2.3, an alternative to early fusion (e.g. feature vector concatenation) is late fusion. Therefore, considering the audio feature $A3$ as the best one for audio-visual combination, we have run a set of experiments aiming at analyzing the results of late fusion for this problem.

For that purpose, we have employed individual classifiers for video and audio, and then, another SVM has been trained on the scores returned by the individual classifiers. Let $s^c_{hog}$, $s^c_{hof}$ and $s^c_{au}$ be the scores returned by SVM trained on HOG, HOF and audio features for category $c$, respectively. We define a new feature vector $\mathbf{f}^c_{lf}$, for a given video, as the concatenation of $s^c_{hog}$, $s^c_{hof}$ and $s^c_{au}$. A new SVM is trained on the new set of features $\mathbf{f}^c_{lf}$.

The results obtained are shown in Tab. 6 following the same rationale employed in Tab. 5. As can be observed, in this case audio-visual late fusion does not makes a clear improvement from the single modalities(see positive differences).

In [41] a new technique is suggested for late fusion using the internal order of the items from each classifier to estimate a better combined order. We have also

Table 5: **Performance of audio-video combination.** Statistical tests showing the improvement of audio-visual features over visual HOG+HOF. First column (i.e. left) indicates both fusion method and kind of audio feature that is combined with HOG+HOF. Columns 2nd to 7th indicate the different sizes of the visual vocabularies tested. Last column contains the mean of the *Succ* values obtained for the given fusion method. Each cell contains $(d, H, m)$, where $d$ is the average mean difference between the classifiers (i.e. positive values indicate that audio-visual features improve over only visual ones), $H$ denotes the most likely hypothesis (i.e. $H_1$ indicates significant difference), and $m$ represents the average on the performance *Succ* for the audio-visual classifiers. All the $m$ values greater than the best value obtained by a single modality (i.e. $> 48.5$) are marked in bold. See text for discussion.

| | HOG+HOF | | | | | | |
|---|---|---|---|---|---|---|---|
| Audio-Visual | K100 | K500 | K1000 | K2000 | K1000-dense | K2000-dense | Mean-AV |
| BLF-*raw* | $(+5.36, H_1, 47.36)$ | $(-1.14, H_0, 44.86)$ | $(+1.00, H_0, 43.50)$ | $(-0.07, H_0, 43.93)$ | $(+0.21, H_0, 44.71)$ | $(-1.29, H_0, 43.71)$ | 46.57 |
| BLF-A1 | $(+6.14, H_1, 48.14)$ | $(-1.07, H_0, 44.93)$ | $(+1.50, H_0, 44.00)$ | $(+1.29, H_0, 45.29)$ | $(+2.50, H_1, 47.00)$ | $(+2.14, H_1, 47.14)$ | |
| BLF-A2 | $(+7.79, H_1, \mathbf{49.79})$ | $(+2.07, H_1, 48.07)$ | $(+0.64, H_0, 43.14)$ | $(+1.43, H_0, 45.43)$ | $(+1.07, H_0, 45.57)$ | $(+3.57, H_1, \mathbf{48.57})$ | |
| BLF-A3 | $(+7.50, H_1, \mathbf{49.50})$ | $(+2.64, H_0, \mathbf{48.64})$ | $(+5.79, H_1, 48.29)$ | $(+3.71, H_1, 47.71)$ | $(+4.50, H_1, \mathbf{49.00})$ | $(+4.50, H_1, \mathbf{49.50})$ | |
| *Mean-Succ-BLF* | 48.67 | 46.66 | 44.73 | 45.59 | 46.57 | 47.23 | |
| MKL-raw | $(+3.00, H_1, 45.00)$ | $(-4.57, H_1, 41.43)$ | $(+0.86, H_1, 43.36)$ | $(-0.21, H_0, 43.79)$ | $(+0.57, H_0, 45.07)$ | $(-0.64, H_0, 44.36)$ | 46.26 |
| MKL-A1 | $(+5.43, H_1, 47.43)$ | $(+2.21, H_1, 48.21)$ | $(+4.93, H_1, 47.43)$ | $(+2.07, H_1, 46.07)$ | $(+5.00, H_1, \mathbf{49.50})$ | $(+2.64, H_1, 47.64)$ | |
| MKL-A2 | $(+5.36, H_1, 47.36)$ | $(+0.50, H_0, 46.50)$ | $(+5.00, H_1, 47.50)$ | $(+3.43, H_1, 47.43)$ | $(+6.36, H_1, \mathbf{50.86})$ | $(+4.71, H_1, \mathbf{49.71})$ | |
| MKL-A3 | $(+1.79, H_0, 43.79)$ | $(+0.43, H_0, 46.43)$ | $(+3.57, H_1, 46.07)$ | $(+3.00, H_1, 47.00)$ | $(-1.57, H_0, 42.93)$ | $(+0.36, H_0, 45.36)$ | |
| *Mean-Succ-MKL* | 45.89 | 45.64 | 46.09 | 46.07 | 47.09 | 46.77 | |
| Bimodal-*raw* | $(+2.86, H_1, 44.86)$ | $(-3.07, H_1, 42.93)$ | $(+2.14, H_0, 44.64)$ | $(-0.29, H_0, 43.71)$ | $(-1.71, H_1, 42.79)$ | $(-1.36, H_0, 43.64)$ | 45.93 |
| Bimodal-A1 | $(+2.93, H_0, 44.93)$ | $(-1.21, H_0, 44.79)$ | $(+0.36, H_0, 42.86)$ | $(+1.79, H_1, 45.79)$ | $(+1.50, H_0, 46.00)$ | $(+2.00, H_0, 47.00)$ | |
| Bimodal-A2 | $(+4.64, H_1, 46.64)$ | $(+0.64, H_0, 46.64)$ | $(+2.79, H_0, 45.29)$ | $(+1.21, H_0, 45.21)$ | $(+3.00, H_1, 47.50)$ | $(+2.79, H_1, 47.79)$ | |
| Bimodal-A3 | $(+6.14, H_1, 48.14)$ | $(+3.07, H_1, \mathbf{49.07})$ | $(+5.93, H_1, 48.43)$ | $(+1.50, H_0, 45.50)$ | $(+5.07, H_1, \mathbf{49.57})$ | $(+3.57, H_0, \mathbf{48.57})$ | |
| *Mean-Succ-Bimodal* | 46.14 | 45.86 | 45.30 | 45.05 | 46.46 | 46.75 | |
| *Summary* | $(+5.01, H_1, 47.01)$ | $(+0.12, H_0, 46.12)$ | $(+2.95, H_1, 45.45)$ | $(+1.58, H_1, 45.58)$ | $(+2.25, H_1, 46.75)$ | $(+1.95, H_1, 46.95)$ | |

Table 6: **Performance of late fusion of audio-visual features using A3 as the audio feature.** The table shows the results of comparing visual with audio-visual features using the late fusion approach. Each cell contains $(d, H, m)$ (see text for details and discussion).

| Audio/HOG+HOF | K100 | K500 | K1000 | K2000 | K1000-dense | K2000-dense |
|---|---|---|---|---|---|---|
| A3 | $(+2.43, H_1, 44.43)$ | $(-1.21, H_0, 44.79)$ | $(-1.50, H_0, 41.00)$ | $(+0.79, H_0, 44.79)$ | $(-4.00, H_1, 40.50)$ | $(-0.93, H_0, 44.07)$ |

tested this technique in our data, however it did not show any improvement on the results from the previous SVM late classifiers.

### 3.5 Discussion

The human interaction categorization (HIR) task has been studied in this paper as a function of its two main modalities, audio and video. The experimental results shown in Tab. 1 and Tab. 3 indicate that decoding of category information from a single modality is still a very difficult task. The single modality best score is obtained from the audio with a 48.5% of sucess, what is a low rate. In our understanding, the video signal is plenty of information but coded in the images in a very complex way, what added to the high number of degree of freedom defining each interaction, makes very hard to decode relevant features. The audio signal is simpler and therefore easier for processing.

The results shown in Tab. 5 indicate that the early fusion approach is an improving strategy for HIR categorization. It is observed that many of the audio-visual combinations improve the best single modality score, pointing out that the feature combination benefits when adequate features are selected (boldface data in Tab. 5). The combination $K1000$-*dense* and $MKL$-$A2$ obtains the best average results for our task. However, it is remarkable that the greatest amount of successful audio-video combinations – that is the combinations with higher score than the best from the single modalities $(48,5\%)$ – and the best full average score are associated to the baseline strategy (see column *Mean-AV* in Tab. 5). For the baseline approach, the combination ($BLF$-$A2$,$K100$) shows the best score with an average score of 49.79, and, for the bimodal approach, the best one is the combination $K1000$-*dense* with *bimodal*-$A3$ with an average score of 49.5. These results show very small differences among the three strategies as on average as in the highest scores. Nevertheless, the baseline strategy shows its best score when the shortest codeword is used to code the video ($K100$), but the other two approaches prefer a large codeword with dense sam-

pling. For the audio, the best features seem to be given by $A2$. All these results show the importance of selecting features according to the classifier to use. If we focus on the row named *Summary* of Tab. 5, we can see that all the differences are positive, what means that, in general, the audio-visual features improve on the visual ones. In addition, we can see in Tab. 8 that the best *Succ* value achieved with audio-visual features (i.e. 54.5) is clearly superior to the best one reported with a single modality (i.e. 48.5 in Tab. 3).

With regard to late fusion, in Tab. 6 it can be observed that this type of late fusion performs worse than the early fusion approach. The results show that in most of the cases, the video features alone obtain better results than the combination (see negative differences). This result could be expected looking at the low classification scores obtained from each single modality. This means very noisy inputs for the late fusion algorithm making very difficult to recognize the true audio-visual patterns.

In order to shed some light on the improvement provided by the audio features on the four evaluated interaction categories, we report in Tab. 7 the results of a statistical study performed on the audio-visual approach that achieved the best mean results on the study presented in Tab. 5: MKL-A2 with HOG+HOF-K1000-dense. We can observe that both *high-five* and *kiss* categories clearly benefits from audio-visual features (i.e. positive differences supported by $H_1$), and *hug* as well but in a moderated manner. In contrast, *hand-shake* does not. Watching the actual video clips of the dataset used in our experiments, we notice that both *high-five* and *kiss* have associated a sound pattern (i.e. kind of brief outburst) very distinctive, at least for humans, unlike *hug* has. For the case of *hand-shake* our impression is that since it does not have always associated a sound pattern as clear as the other two commented interactions have, the few greeting words that are sometimes said during the interaction introduce uncertainty in the system.

*Comparison with the state-of-the-art* In addition to *Succ*, and for comparison purposes with [20], we compute the mean Average Precision (mAP) – as in a video retrieval setup – as follows: *(i)* we train the models (one model per positive class, i.e. 4 models=SVM) with subset A and we classify samples on subset B; *(ii)* we train the models (one model per positive class, i.e. 4 models=SVM) with subset B and we classify samples on subset A; and, *(iii)* we put together all the classified samples from *(i)* and *(ii)* in the same *bucket*, along with their corresponding scores, in order to compute a *global* Precision-Recall curve; and, *(iv)* the area under the precision-recall curve (AP) is used as performance measurement for each class. Note that, since there are 4 positive classes, we compute one AP at a time following the previously explained procedure. Finally, the average AP over the 4 classes is reported as mAP. Note that all the negative videos are included in this evaluation (i.e. video retrieval task).

Tab. 8 presents a selection of the two best results achieved in our experiments for the evaluated audio-visual features (see Tab. 5). The results shown in column *Succ(4)* correspond to measurement *Succ* applied over the four categories of interactions, as done in the previous sections. However, column *Succ(4+neg)* includes the *negative* samples of TVHID as an additional fifth category.

Column *mAP(4+neg)* in Tab. 8 allows a direct comparison with the state-of-the-art on video retrieval on TVHID. The best configuration found for our audio-visual proposal (i.e. 0.4779) is around 13% better than the one reported by Patron-Perez *et al.* [20] with their fully automatic setup (i.e. 0.4244). Although this mAP is still below the 0.5074 achieved in [20] when *manual* tracks of persons are used as input. Note that in our experiments we only use the location of the persons during training, to learn a dictionary of *clean* STIPs.

*Recommendations* From the results obtained in our experimental evaluation, our recommendations for fusing visual and audio information for the task of HIR are: (i) non-dense HOG+HOF for visual features ; (ii) MFCC for audio features; (iii) early fusion instead of late fusion; (iv) MKL as fusion scheme due to the equivalence between MKL-linear and BLF. BLF and MKL show similar mean-AV (see Tab.5) meaning that, in this problem, the used descriptors have a similar and additive contribution. In this way, BLF does not require an additional learning step as MKL does (i.e. kernel combination weights), therefore, our first choice would be BLF (i.e. simple concatenation of feature vectors). Nevertheless, the best strategy would be to estimate the **d**-parameters using a MKL-linear model; (v) for visual features, a large dictionary size (i.e. around 1000 words) leads, in general, to better mean performance, regardless the size of the audio dictionary; (vi) however, small or medium sized audio dictionaries (i.e. around 100 words) are preferred.

## 4 Conclusions

In this paper, we have presented a new focus on the problem of human interaction categorization in TV Videos. In contrast to other common approaches in the field of human action/interaction recognition, we show in this

Table 7: **Performance of audio-video combination for each class in the test *MKL-A2* on *K1000-dense*.** Statistical tests showing the improvement of audio-visual features over visual ones on each interaction category: hug (HU), kiss (KI), hand-shake (HS), high-five (HF). Each cell contains $(d, H, m)$ (see text for details and discussion).

| *HU* | *KI* | *HS* | *HF* |
|---|---|---|---|
| $(+0.86, H_0, 66.86)$ | $(+8.86, H_1, 40.86)$ | $(-7.71, H_1, 36.29)$ | $(+23.43, H_1, 59.43)$ |

Table 8: **Summary of the best results on TVHID.** Percentage of correct categorization *Succ* and mAP. Column *Succ(4+neg)* and *mAP(4+neg)* are included for comparison purposes (see [20]). Column *Succ(4+neg)* includes the *negative* samples as an additional fifth category.

| Feats/Perf | *Succ(4)* | *Succ(4+neg)* | *mAP(4+neg)* |
|---|---|---|---|
| BLF-HOGHOF-K100+*A*2-K200 | 54.5 | **44.7** | **0.4779** |
| MKL-HOGHOF-K500+*A*3-K25 | 54.5 | 39.3 | 0.4536 |
| Patron-Perez *et al.* [20] | N/A | 40.4 | 0.4244 |

paper (i) that human interaction categorization is a problem better defined by audio-visual information; (ii) that each single modality (audio or video) contains too much uncertainty to achieve good categorization scores by itself; (iii) that the audio, as a single modality, is simple to process providing, on average, more discriminative features for HIR than the image-based ones, moreover, audio by itself also provides higher score than the average of audio-video combinations; (iv) that the combination of audio and visual features, when successful, makes a significant improvement on the categorization score in comparison with the single modalities; (v) that the size of the coding dictionary for the visual signal appears a relevant factor for the combination strategy; and, (vi) that the audio-visual framework offers promising results in comparison with the state-of-the-art on TVHID, in terms of mean average precision.

In conclusion, the results of this work confirm that human interaction categorization is a matter of audio-visual features combination where the selected features and the way we combine them are relevant steps in order to improve the final performance.

In addition, we think that the addition of a voice recognition stage could help significantly for identifying some interactions where people typically speak as for example *hand-shake*. This will be a line of future research.

## Acknowledgements

This research was partially supported by the Spanish Ministry of Economy and Competitiveness under projects P10-TIC-6762, TIN2010-15137 and TIN2012-32952 and the European Regional Development Fund (FEDER)

## References

1. Bakker, E., Lew, M.: Semantic video retrieval using audio analysis. In: M. Lew, N. Sebe, J. Eakins (eds.) Image and Video Retrieval, *Lecture Notes in Computer Science*, vol. 2383, pp. 201–218 (2002)
2. Bredin, H., Koenig, L., Farinas, J.: Irit at trecvid 2010: Hidden markov models for context-aware late fusion of multiple audio classifiers. In: TRECVID 2010 Notebook papers (2010)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE Computer Society, Washington, DC, USA (2005)
4. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision (2006)
5. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on **28**(4), 357 – 366 (1980)
6. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: Advances in Neural Information Processing Systems (2011)
7. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007)
8. Inoue, N., Wada, T., Kamishima, Y., Shinoda, K., Sato, S.: Tokyotech+canon at trecvid 2011. In: TRECVID 2011 Notebook papers (2011)
9. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings of ACM International Conference on Multimedia Retrieval (ICMR), oral session (2011)
10. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64(2/3)**, 107–123 (2005)
11. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
12. Laptev, I., Pérez, P.: Retrieving actions in movies. In: Proceedings of the International Conference on Computer Vision, pp. 1–8 (2007)

13. Lartillot, O., Toiviainen, P.: MIR in Matlab (ii): A toolbox for musical feature extraction from audio. In: ISMIR, pp. 127–130 (2007)
14. Li, Y., Mou, L., Jiang, M., Su, C., Fang, X., Qian, M., Tian, Y., Wang, Y., Huang, T., Gao, W.: Pku-idm at trecvid 2010: Copy detection with visual-audio feature fusion and sequential pyramid matching. In: TRECVID 2010 Notebook papers (2010)
15. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: L.M.L. Cam, J. Neyman (eds.) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
16. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics **18**(1), 50–60 (1947)
17. Marin-Jimenez, M., Zisserman, A., Ferrari, V.: "Here´s looking at you, kid". Detecting people looking at each other in videos. In: Proceedings of the British Machine Vision Conference (2011)
18. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009)
19. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(3), 305Ũ–317 (2005)
20. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in tv shows. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(12), 2441–2453 (2012)
21. Perera, A.G.A., Oh, S., Leotta, M., Kim, I., Byun, B., Lee, C.H., McCloskey, S., Liu, J., Miller, B., Huang, Z.F., Vahdat, A., Yang, W., Mori, G., Tang, K., Koller, D., Fei-Fei, L., Li, K., Chen, G., Corso, J., Fu, Y., Srihari, R.: Genie trecvid2011 multimedia event detection: Late-fusion approaches to combine multiple audio-visual features. In: TRECVID 2011 Notebook papers (2011)
22. Press, W.H., Teukolsky, S.A., Vetterling, W., Flannery, B.P.: Numerical Recipes in C++: The Art of Scientific Computing, 2 edn. Cambridge University Press (2002)
23. Reid, I., Benfold, B., Patron, A., Sommerlade, E.: Understanding interactions and guiding visual surveillance by tracking attention. In: R. Koch, F. Huang (eds.) Computer Vision Ũ ACCV 2010 Workshops, *Lecture Notes in Computer Science*, vol. 6468, pp. 380–389. Springer Berlin / Heidelberg (2011)
24. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: Proceedings of the International Conference on Computer Vision, pp. 1593–1600 (2009)
25. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Proceedings of the International Conference on Pattern Recognition, vol. 3, pp. 32–36. Cambridge, U.K. (2004)
26. Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: On the use of audio events for improving video scene segmentation. In: Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on, pp. 1 –4 (2010)
27. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1470–1477 (2003)
28. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York, NY, USA (2006)
29. Stevens, S., Volkmann, J., Newman, E.: A scale for the measurement of the psychological magnitude of pitch. J. Acoust Soc Amer **8**, 185–190 (1937)
30. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. Circuits and Systems for Video Technology, IEEE Transactions on **18**(11), 1473–1488 (2008)
31. Tzanetakis, G., Chen, M.: Building audio classification for broadcast news retrieval. In: Proc. of WIAMIS (2004)
32. Varma, M., Babu, B.R.: More generality in efficient multiple kernel learning. In: ICML, p. 134 (2009)
33. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)
34. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 480–492 (2012)
35. Vishwanathan, S.V.N., Sun, Z., Theera-Ampornpunt, N., Varma, M.: Multiple kernel learning and the SMO algorithm. In: Advances in Neural Information Processing Systems (2010)
36. Wang, H., Ullah, M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings of the British Machine Vision Conference, p. 127 (2009)
37. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding (2006)
38. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)
39. Yao, B., Jiang, X., Khosla, A., Lin, A., Guibas, L., Fei-Fei, L.: Action recognition by learning bases of action attributes and parts. In: Proceedings of the International Conference on Computer Vision. Barcelona, Spain (2011)
40. Ye, G., Jhuo, I.H., Liu, D., Jiang, Y.G., Lee, D.T., Chang, S.F.: Joint audio-visual bi-modal codewords for video event detection. In: ICMR, p. 39 (2012)
41. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F.: Robust late fusion with rank minimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3021–3028 (2012)