Identifying Human Behaviors Using Synchronized Audio-Visual Cues

Michalis Vrigkas, *Student Member, IEEE,* Christophoros Nikou, *Senior Member, IEEE,*
and Ioannis A. Kakadiaris, *Senior Member, IEEE*

Abstract—In this paper, a human behavior recognition method using multimodal features is presented. We focus on modeling individual and social behaviors of a subject (e.g., friendly/aggressive or hugging/kissing behaviors) with a hidden conditional random field (HCRF) in a supervised framework. Each video is represented by a vector of spatio-temporal visual features (STIP, head orientation and proxemic features) along with audio features (MFCCs). We propose a feature pruning method for removing irrelevant and redundant features based on the spatio-temporal neighborhood of each feature in a video sequence. The proposed framework assumes that human movements are highly correlated with sound emissions. For this reason, canonical correlation analysis (CCA) is employed to find correlation between the audio and video features prior to fusion. The experimental results, performed in two human behavior recognition datasets including political speeches and human interactions from TV shows, attest the advantages of the proposed method compared with several baseline and alternative human behavior recognition methods.

Index Terms—Hidden conditional random fields, audio-visual synchronization, multimodal fusion, canonical correlation analysis, human behavior recognition.

1 Introduction

Recognizing human behaviors from video sequences is a challenging task [1], [2]. A behavior recognition system may provide information about the personality and psychological state of a person. Its applications vary from video surveillance to human-computer interaction. Human behavior is often expressed as a combination of non-verbal multimodal cues such as gestures, facial expressions and auditory cues. The correlation between cues from different modalities has been shown to improve recognition accuracy [3]–[5].

The problem of human behavior recognition is challenging for several reasons. First, constructing a visual model for learning and analyzing human movements is difficult. Second, the fine differences between similar classes and the short time duration of human movements make the problem difficult to address. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. The variation of appearance, lighting conditions and frame resolution makes the recognition problem amply challenging. Finally, the inadequate benchmark datasets pose a challenge.

When attempting to recognize human behaviors, one must determine the kinematic states of a person. From psychological point of view, human behaviors may be classified in three types: behavioral, cognitive and social [6]. Our goal is to understand not only social behaviors (e.g., relationships and interactions between people such as hugging or kissing) but also individual behaviors (e.g., expression of personal feelings such as aggressiveness or friendliness).

Factors that can affect human behavior may be decomposed into several components including emotions, moods, actions and interactions with other people. Hence, the recognition of complex actions may be crucial for understanding human behavior. Recognizing human actions that correspond to a specific emotional state of a person or an affective label such as boredom, or kindness, may help understand social behaviors. The task of learning human behaviors is to identify the psychological state or the social activities of a person taking place in the surroundings [7]. Several affective computing methods [8], [9] used semantic annotations in terms of arousal and valence to capture the underlying affect from multimodal data. However, obtaining affective labels for real world data is a challenging task [10] and it may lead to biased representation of human behaviors.

The dimensionality of audio and visual data poses significant challenges to audio-visual analysis. Video features are much more complex and high dimensional than audio, and thus techniques for dimensionality reduction play an important role [11]. In the literature, there are two main fusion strategies, which can be used to tackle this problem [3], [12]. The *early fusion* or fusion at the feature level, combines features from different modalities, usually by reducing the dimensionality of features from each modality and creating a new feature vector that represents the individual. Canonical correlation analysis

- *M. Vrigkas and C. Nikou are with the Department of Computer Science and Engineering, University of Ioannina, Ioannina GR 451 10, Greece.*
  E-mail: mvrigkas@cs.uoi.gr; cnikou@cs.uoi.gr
- *I. A. Kakadiaris is with the Computational Biomedicine Lab, Department of Computer Science, University of Houston, 4800 Calhoun Rd, Houston, TX 77204, USA.*
  E-mail: ioannisk@uh.edu

(CCA) [13] was widely studied in the literature as an effective way for fusing data at feature level [14], [15]. The advantage of early fusion is that it yields good recognition results when the different modalities are highly correlated, since only one learning phase is required. On the other hand, the difficulty of combining the different modalities may lead to the domination of the strongest modality.

The second category of methods, which is known as *late fusion* or fusion at the decision level, combines several probabilistic models to learn the parameters of each modality separately. Then all scores are combined together in a supervised framework yielding a final decision score [16]. The individual strength of each modality may lead to better recognition results. However, this strategy is time consuming and requires more complex supervised learning methods, which may cause a potential loss of the inter-modality correlation. A comparison of early versus late fusion methods for video analysis was reported by Snoek *et al.* [17].

In this work, we address the problem of multimodal data association for human behavior recognition. First, audio and visual data from the video sequences are extracted and then a feature pruning technique is applied to remove redundant features according to the spatiotemporal neighborhood of the features in the video frames. Then, CCA [13] is employed to find the synchronization offset between the audio and video features, such that the correlation between sound emissions and human movements is maximized. Finally, the projected data are concatenated into a new feature vector and are used as input to a chain hidden conditional random field (HCRF) [18] model to capture the interaction across modalities and compute the underlying hidden dynamics between the labels and the features. Our method is also able to cope with videos with varying human poses as feature pruning may reduce the background and discard irrelevant frames. In contrast to most of the multimodal human behavior analysis methods, the combination of feature pruning and early fusion keeps the complexity of our method relatively low, as only one step of classification for estimating human behaviors is required.

The contributions of this paper can be summarized as follows:

- We developed a supervised multimodal learning framework, for human behavior recognition based on the canonical correlation of audio and visual features.
- We proposed a feature selection technique for pruning redundant features, based on the spatiotemporal neighborhood of the visual features that reduced the complexity of the classification algorithm.
- We employed an audio-visual synchronization method to temporally align the audio and video features, to better exploit the correlation of the audio-visual features and improve the recognition accuracy.
- We introduced a novel behavior dataset, called the *Parliament* dataset [19] and conducted comprehensive experiments to assess the effect of the audio information on the behavioral recognition task.

Although the *Parliament* dataset was first introduced by Vrigkas *et al.* [19], it is in this paper that audio information is employed to enhance the recognition accuracy for this dataset. The main difference with respect to [19], is that in [19] a fully connected conditional random field (CRF) [20] model is employed, where different labels for each video frame were considered. This makes the model more suitable to handle video sequences with more than one label per video, but it significantly increases the complexity of the model.

To evaluate our method, we used two publicly available datasets, the *Parliament* dataset [19] with three behavioral labels: *friendly*, *aggressive*, and *neutral* and the TV human interaction (TVHI) dataset [21], which contains four different interaction activities: *hand shakes*, *high fives*, *hugs* and *kisses*.

The remainder of the paper is organized as follows: in Section 2, a brief review of the related work is presented. Section 3 presents the proposed approach including the feature selection method and the audio-visual synchronization technique. In Section 4, the novel *Parliament* dataset is presented and experimental results are reported. Finally, conclusions are drawn in Section 5.

2 Related Work

In this paper, the term "behavior" is used to describe both activities and events, which are captured in a video sequence. We categorize the human behavior recognition methods into two main categories: unimodal and multimodal. The latter is of great interest, as several multimodal fusion techniques have been widely studied in the literature.

2.1 Unimodal Behavior Recognition Methods

Much research has focused on unimodal behavior recognition methods. Social interactions are an important part of human daily life. Fathi *et al.* [22] modeled social interactions by estimating the location and orientation of the faces of the persons taking part in a social event, computing a line of sight for each face. This information is used to infer the location an individual person attended. The type of interaction is recognized by assigning social roles to each person. The authors were able to recognize three types of social interactions: dialogue, discussion and monologue. Ramanathan *et al.* [23] aimed at assigning social roles to people associated with an event. They formulated the problem by using a CRF [20] model to describe the interactions between people. Tran *et al.* [24] presented a graph-based clustering algorithm to discover interactions between groups of people in a crowd scene. A bag-of-words approach was used to describe the group activity, while an SVM classifier was used to recognize the human activity. An advantage of CRF-based methods is that they can model arbitrary features of observation sequences.

The problem of multi-person interactions is presented by Burgos *et al.* [25], where the social behavior of mice is discussed. Each video sequence is segmented into periods of activities by constructing a temporal context that combines spatio-temporal features. Morency *et al.* [26] first introduced the latent-dynamic conditional random field (LDCRF) for gesture recognition. They used hidden states to model the sub-structure of each class and learn the dynamics between the class labels. The main difference between the LDCRF model and the HCRF [18] is that the former contains a class label per observation, which makes it suitable for recognizing unsegmented sequences.

Patron-Perez *et al.* [21] introduced a method for recognizing dyadic human interactions in TV shows by tracking a person through time and using head pose orientations for extracting useful information about the interactions. Gaidon *et al.* [27] addressed the problem of human action recognition by introducing a supervised method for clustering motion trajectories and representing a hierarchical scheme for long human actions. Li *et al.* [28] have also used trajectories to tackle the problem of human action recognition using canonical correlation to better exploit the intra-class variations of data. In general, although these methods may perform well under some circumstances, they suffer from the problem of data association. That is, these methods are based on the collection of time series of spatio-temporal features at single pixel locations. However, the same pixel location does not represent the same information over time as acting humans are considered as highly deformable objects. Thus, collecting time series may require tracking of visual features in time.

2.2 Multimodal Behavior Recognition Methods

Recently, much attention has been focused on multimodal behavior recognition methods. An event can be described by different types of features that provide more and useful information. In this context, several multimodal methods are based on feature fusion, which can be expressed by two different strategies: early fusion and late fusion. The easiest way of gaining the benefits of multiple features is to directly concatenate features in a larger feature vector and then learn the underlying action [29]. This feature fusion technique may improve recognition performance, but the new feature vector is of much larger dimension.

Audio-visual representation of human actions has gained an important role in human behavior recognition methods. Marín-Jiménez *et al.* [30] used a bag of visual-audio words scheme along with late fusion technique for recognizing human interactions in TV shows. Even though their method performs well in recognizing human interaction, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem. Bousmalis *et al.* [5] considered a system based on HCRFs [18] for spontaneous agreement and disagreement recognition using audio and visual features. Wang *et al.* [31] proposed a semi-supervised framework for recognizing human actions combining different visual features. Although both methods yielded promising results, they did not consider any kind of explicit correlation and/or association between the different modalities.

Sargin *et al.* [32] suggested a method for speaker identification integrating a hybrid scheme of early and late fusion of audio-visual features and used CCA [13] to synchronize the multimodal features. However, their method can cope with video sequences of frontal view only. Wu *et al.* [33] proposed a human activity recognition system by taking advantage of the auditory information of the video sequences of the HOHA dataset [34] and used late fusion techniques for combining audio and visual cues. The main disadvantage of this method is that it used different classifiers to separately learn the audio and visual context. Also, the audio information of the HOHA dataset contains dynamic backgrounds and the audio signal is highly diverse (i.e., audio shifts roughly from one event to another), which creates the need for developing audio features selection techniques. Similar in spirit is the work of Wu *et al.* [35], who used the generalized multiple kernel learning algorithm for estimating the most informative audio features, while they applied fuzzy integral techniques to combine the outputs of two different SVM classifiers increasing the computational burden of the method.

Song *et al.* [4] proposed a novel method for human behavior recognition based on multi-view hidden conditional random fields (MV-HCRF) [36] and estimated the interaction of the different modalities by using kernel canonical correlation analysis (KCCA) [13]. However, their method cannot address the challenge of data that contain complex backgrounds, and due to the downsampling of the original data the audio-visual synchronization may be lost. Also, their method used different sets of hidden states for audio and visual information. This property considers that the audio and visual features were a priori synchronized, while it increases the complexity of the model. Siddique *et al.* [37] analyzed four different affective dimensions such as activation, expectancy, power and valence [38]. To this end, they proposed joint hidden conditional random Fields (JHCRF) as a new classification scheme to take advantage of the multimodal data. Furthermore, their method uses late fusion to combine audio and visual information together. This may lead to significant loss of the inter-modality dependence, while it suffers from carrying the classification error across different levels of classifiers. Although their method could efficiently recognize the affective state of a person, the computational burden was high because JHCRFs require twice as many hidden variables as the traditional HCRFs when features represent two different modalities.

An audio-visual analysis for recognizing dyadic interactions was presented by Yang *et al.* [39]. The author combined a Gaussian Mixture Model (GMM) [40] with a Fisher kernel to model multimodal dyadic interactions and predict the body language of each subject according to the behavioral state of his/her interlocutor. Castellano *et al.* [41] explored the dynamics of body movements to identify affective behaviors using time series of multimodal data. Martinez *et al.* [42] presented a detailed review of learning methods for classification of affective and cognitive states of computer game players. They analyzed the properties of directly using affect annotations in classification models, and proposed a method for transforming such annotations to build more accurate models. Nicolaou *et al.* [43] proposed a regression model based on support vector machines for regression (SVR) for continuous prediction of multimodal emotional states, using facial expression, shoulder gesture, and audio cues in terms of arousal and valence.

Multimodal affect recognition methods in the context of neural networks and deep learning have generated considerable recent research interest [44]. Metallinou *et al.* [45] employed several hierarchical classification models from neural networks to hidden Markov models and their combinations to recognize audio-visual emotional levels of valence and arousal rather than emotional labels such as anger or kindness. Kim *et al.* [46] used deep belief networks (DBN) [47] in both supervised and unsupervised manner to learn the most informative audio-visual features and classify human emotions in dyadic interactions. Their system was able to preserve non-linear relationships between multimodal features and shown that unsupervised learning can be used efficiently for feature selection. In a more recent study Martinez *et al.* [48] could efficiently extract and select the most informative multimodal features using deep learning, to model emotional expressions and recognize the affective states of a person. They incorporated psychological signals into emotional states such as relaxation, anxiety, excitement and fun, and demonstrated that deep learning was able to extract more informative features than feature extraction on psychological signals.

3 The Proposed Approach

We assume that a set of training labels is provided and each video sequence is pre-processed to obtain a bounding box of the human in every frame and each person is associated with a behavioral label. The model is general and can be applied to several behavior recognition datasets. Our method uses HCRFs, which are defined as a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 1), as the probabilistic framework for modeling the behavior of a subject in a video. First, audio and visual features are computed in each video frame capturing the roles associated with the bounding boxes. Next, irrelevant visual features are eliminated according to their spatio-temporal relationship of neighboring features. Then, the synchronization offset between the different modalities is estimated by using CCA. Finally, belief propagation (BP) [49] is applied to estimate the labels.

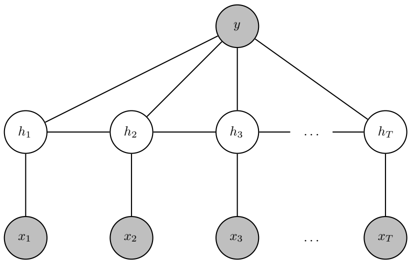


Fig. 1: Graphical representation of the chain structure model. The grey nodes are the observed features and the unknown labels represented by x and y, respectively. The white nodes are the unobserved hidden variables h.

3.1 Multimodal HCRF

We consider a labeled dataset with N video sequences $\mathcal{D} = \{\mathbf{x}_{i,j}, y_i\}_{i=1}^{N}$, where $\mathbf{x}_{i,j} = (\mathbf{a}_{i,j}, \mathbf{v}_{i,j})$ is a multimodal observation sequence, which contains audio ($\mathbf{a}_{i,j} \in \mathbb{R}^{n_a \times T}$) and visual data ($\mathbf{v}_{i,j} \in \mathbb{R}^{n_v \times T}$) of length T with $j = 1 \dots T$. For example, $\mathbf{x}_{i,j}$ corresponds to the j^{th} frame of the i^{th} video sequence. Finally, y_i corresponds to a class label defined in a finite label set $\mathcal{Y}$. Our model is applied to all video sequences in the training set. In what follows, we omit indices i and j for simplicity.

It is useful to note that our HCRF model is a member of the exponential family and the probability of the class label given an observation sequence is given by:

$$\begin{aligned} p(y|\mathbf{x};\mathbf{w}) &= \sum_{\mathbf{h}} p(y,\mathbf{h}|\mathbf{x};\mathbf{w}) \\ &= \sum_{\mathbf{h}} \exp\left(E(y,\mathbf{h}|\mathbf{x};\mathbf{w}) - A(\mathbf{w})\right) , \end{aligned} \tag{1}$$

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters, $\mathbf{h} = \{h_1, h_2, \dots, h_T\}$, with $h_i \in \mathcal{H}$ is a set of latent variables. In particular, the number of latent variables may be different from the number of samples, as h_j may correspond to a substructure in a sample. However, for simplicity we use the same notation. Finally, $E(y,\mathbf{h}|\mathbf{x};\mathbf{w})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = \log \sum_{y'} \sum_{\mathbf{h}} \exp\left(E(y',\mathbf{h}|\mathbf{x};\mathbf{w})\right) . \tag{2}$$

Different sufficient statistics $E(y,\mathbf{h}|\mathbf{x};\mathbf{w})$ in (1) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$\begin{aligned} E(y,\mathbf{h}|\mathbf{x};\mathbf{w}) &= \sum_{j \in \mathcal{V}} \sum_{\ell} \Phi_\ell(y, h_j, \mathbf{x}; \boldsymbol{\theta}_\ell) \\ &\quad + \sum_{j,k \in \mathcal{E}} \sum_{\ell} \Psi_\ell(y, h_j, h_k; \boldsymbol{\omega}_\ell) , \end{aligned} \tag{3}$$

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned and $\Phi_\ell(y, h_j, \mathbf{x}; \boldsymbol{\theta}_\ell)$, $\Psi_\ell(y, h_j, h_k; \boldsymbol{\omega}_\ell)$ are the unary and pairwise potentials, respectively.

The unary potential is expressed by:

$$\Phi_\ell(y, h_j, \mathbf{x}; \boldsymbol{\theta}_\ell) = \sum_j \phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) + \sum_j \phi_{2,\ell}(h_j, \mathbf{x}; \boldsymbol{\theta}_{2,\ell}) \,, \tag{4}$$

and it can be considered as a state function, which consists of two different feature functions. The label feature function, which models the relationship between the label y and the hidden variables h_j, is expressed by:

$$\phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) = \sum_{\lambda \in \mathcal{Y}} \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) \,, \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if its argument is true and 0 otherwise. The observation feature function, which models the relationship between the hidden variables h_j and the observations $\mathbf{x}$, defined by:

$$\phi_{2,\ell}(h_j, \mathbf{x}; \boldsymbol{\theta}_{2,\ell}) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{2,\ell} \mathbb{1}(h_j = a) \mathbf{x} \,. \tag{6}$$

The pairwise potential is a transition function and represents the association between a pair of connected hidden states h_j and h_k and the label y. It is expressed by:

$$\Psi_\ell(y, h_j, h_k; \boldsymbol{\omega}_\ell) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a,b \in \mathcal{H}}} \boldsymbol{\omega}_\ell \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) \mathbb{1}(h_k = b) \,. \tag{7}$$

3.2 Parameter Learning and Inference

Our goal is to assign a test video sequence with a behavioral role by maximizing the posterior probability:

$$y = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}; \mathbf{w}) \,. \tag{8}$$

In the training step the optimal parameters $\mathbf{w}^*$ are estimated by maximizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w}) - \frac{1}{2\sigma^2} \|\mathbf{w}\|^2 \,. \tag{9}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}; \mathbf{w})$ and quantifies how well the distribution in (1) defined by the parameter vector $\mathbf{w}$ matches the labels y. It can be rewritten as:

$$\begin{aligned} \log p(y_i|\mathbf{x}_i; \mathbf{w}) = \log &\sum_{\mathbf{h}} \exp(E(y, \mathbf{h}|\mathbf{x}; \mathbf{w})) \\ - \log &\sum_{y', \mathbf{h}} \exp(E(y', \mathbf{h}|\mathbf{x}; \mathbf{w})) \,. \end{aligned} \tag{10}$$

The second term is a Gaussian prior with variance σ^2 and works as a regularizer. The loss function is minimized using a gradient-descent optimization method. More specifically, in our experiments we used the limited-memory BFGS (LBFGS) method to maximize the negative log-likelihood of the data.

Having set the parameters $\mathbf{w}$, the marginal probability is obtained by applying the BP algorithm [40] using the graphical model as depicted in Fig. 1.

3.3 Multimodal Feature Extraction

In this work, we used three different sets of visual features (i.e., STIPs, head orientations, and proxemic features). First, we extract local space-time features at frame rate of 25 fps using a 72-dimensional vector of HoG and 90-dimensional vector of HoF feature descriptors [50] for each STIP [51], which captures the human motion between frames. These features were selected because they can capture salient visual motion patterns in an efficient and compact way.

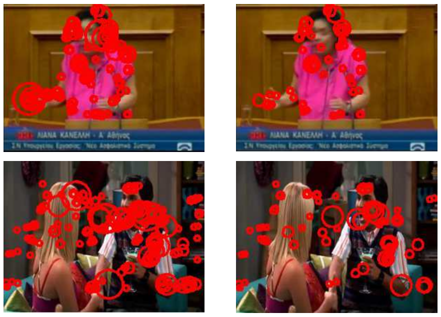
Feature extraction may be erroneous due to cluttered backgrounds caused by camera motion or changes in illumination and appearance. Reducing the number of irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields a more general concept. For this reason, we adopt a similar technique with Liu *et al.* [52] and we perform feature pruning based on spatial and temporal neighborhood of motion features. The proposed algorithm depends on two factors: (i) the distance between the centers of the feature locations and (ii) the scatter of each feature group in consecutive frames.

Let N_t be the number of features in frame t and N be the total number of features in the video sequence. Let also, μ_t and σ_t^2 be the center and the variance of the feature locations in frame t, respectively. First, we discard those frames where N_t is much larger than the mean number of features in the video sequence. Next, if the ratio of the difference of the means to the standard deviation of feature locations and the number of features between frame t and its neighboring frames $t-1$ and $t+1$ are over a predefined threshold, we select $M_t \leq N_t$ features that lie close to the centers of the feature locations in neighboring frames. A detailed description of the proposed feature pruning algorithm is presented in Algorithm 1. Figure 2 depicts some representative examples of the feature pruning technique. Feature pruning may significantly reduce the number of features (Fig. 6).

In cases where the video sequences are not person-centric, but may contain human interactions (e.g., hugging), STIP features are not adequate. For this reason, we have used head orientation as additional feature. This choice is motivated by the fact that a person who interacts with another is more likely to look at that person than looking at somewhere else. Furthermore, we have also used proxemic features, which capture the spatial and temporal relations between interacting persons detected in the video sequences. This means that interacting persons are in general more probable to lie close to each other (spatially and temporally).

Algorithm 1 Feature pruning

```
Input: Original features $\mathbf{v}_t$ for frame $t$.
Output: Pruned features $\mathbf{z}_t$ for frame $t$.
if $N_t >> \text{mean}(N)$ then
    Discard frame $t$;
end if
if $\left(\frac{\|\mu_{t-1} - \mu_t\|^2}{\sigma^2_{t-1} + \sigma^2_t} > \varepsilon \;\&\; \frac{\|\mu_t - \mu_{t+1}\|^2}{\sigma^2_t + \sigma^2_{t+1}} > \varepsilon\right)$ & $(|N_{t-1} - N_t| > \zeta \;\&\; |N_t - N_{t+1}| > \zeta)$ then
    $j \leftarrow 1$;
    for $i \leftarrow 1$ to $N_t$ do
        if $\frac{\|\mu_{t-1} - \mu_t\|^2}{\|\mathbf{v}_{i,t} - \mu_t\|^2} < T \;\&\; \frac{\|\mu_t - \mu_{t+1}\|^2}{\|\mathbf{v}_{i,t} - \mu_t\|^2} < T$ then
            $\mathbf{z}_{j,t} \leftarrow \mathbf{v}_{i,t}$;
            $j \leftarrow j + 1$;
        end if
    end for
end if
```



(a) (b)

Fig. 2: Representative examples of feature pruning. (a) The original features and (b) the pruned features for the *Parliament* dataset [19] (top row) and the TV human interaction dataset [21] (bottom row). Feature pruning may reduce the number of features by 29% on average.

TABLE 1: Types of audio and visual features used for human behavior recognition. The numbers in parentheses indicate the dimension of the features.

Audio features (39)	**Visual features** (166)
MFCCs (13)	STIP (162)
Delta-MFCCs (13)	Head orientations (2)
Delta-delta-MFCCs (13)	Proxemic (2)

Moreover, many audio features have been studied for speaker detection and voice recognition [53]. Mel-frequency cepstral coefficients (MFCCs) [54] are the most popular and common audio features. We employ the MFCCs features and their first and second order derivatives (delta and delta-delta MFCCs) to form an audio feature vector of dimension 39. Table 1 summarizes all audio and visual feature types used in our algorithm.

3.4 Audio-Visual Synchronization and Fusion

The purpose of the proposed method is to perform multimodal human behavior recognition by taking into account both visual and audio information. One drawback of combining features of different modalities is the different frame rate that each modality may have. Thus, prior to the fusion step, visual features are interpolated to match the audio frame rate. However, interpolation may harm the synchronization between the audio and visual features, which is necessary to better exploit the correlation between the different modalities. To this end, we propose using CCA to estimate audio-visual synchronization offset and perform the data fusion.

Given a set of zero-mean paired observations $\{(\mathbf{a}_i, \mathbf{v}_i)\}_{i=1}^{M}$, with $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$, CCA seeks to find two linear transformation vectors $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_v$, such that the correlation $\rho(\boldsymbol{\gamma}_a^T \mathbf{A}, \boldsymbol{\gamma}_v^T \mathbf{V})$ between the projections onto these vectors, $\mathbf{a} = \boldsymbol{\gamma}_a^T \mathbf{A}$ and $\mathbf{v} = \boldsymbol{\gamma}_v^T \mathbf{V}$ (also known as canonical variates) is maximized:

$$\begin{aligned} \rho(\mathbf{a}, \mathbf{v}) &= \max_{\gamma_a, \gamma_v} \frac{\mathbb{E}[av]}{\sqrt{\mathbb{E}[a]^2 \mathbb{E}[v]^2}} \\ &= \max_{\gamma_a, \gamma_v} \frac{\mathbb{E}[\boldsymbol{\gamma}_a^T \mathbf{A} \mathbf{V}^T \boldsymbol{\gamma}_v]}{\sqrt{\mathbb{E}[\boldsymbol{\gamma}_a^T \mathbf{A} \mathbf{A}^T \boldsymbol{\gamma}_a] \mathbb{E}[\boldsymbol{\gamma}_v^T \mathbf{V} \mathbf{V}^T \boldsymbol{\gamma}_v]}} \\ &= \max_{\gamma_a, \gamma_v} \frac{\boldsymbol{\gamma}_a^T \Sigma_{av} \boldsymbol{\gamma}_v}{\sqrt{\boldsymbol{\gamma}_a^T \Sigma_{aa} \boldsymbol{\gamma}_a \mathbf{w}_v^T \Sigma_{vv} \boldsymbol{\gamma}_v}}, \end{aligned} \tag{11}$$

where $\mathbb{E}[\cdot]$ is the expected value, $\Sigma_{aa} \in \mathbb{R}^{n_a \times n_a}$ and $\Sigma_{vv} \in \mathbb{R}^{n_v \times n_v}$ are the covariance matrices, respectively, and $\Sigma_{av} \in \mathbb{R}^{n_a \times n_v}$ is the cross-covariance matrix of $\mathbf{A}$ and $\mathbf{V}$.

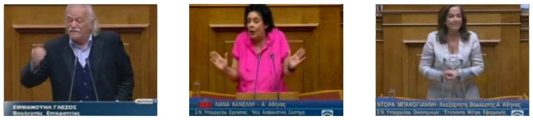
The solutions for $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_v$ are the eigenvectors corresponding to the largest eigenvalues of $\Sigma_{aa}^{-1} \Sigma_{av} \Sigma_{vv}^{-1} \Sigma_{va}$ and $\Sigma_{vv}^{-1} \Sigma_{va} \Sigma_{aa}^{-1} \Sigma_{av}$, respectively.

The greatest challenge when dealing with audio-visual features is to correctly identify the auditory information that corresponds to the motion of the underlying event. This means, that audio and visual features need to be precisely correlated before data fusion is applied [11], [32]. To this end, we assume that there is a time gap τ, which can be seen as an integer offset of frames between audio and visual streams such that the visual feature vector $\mathbf{v}_t$ in frame t corresponds to the $(t + \tau)^{\text{th}}$ audio feature vector $\mathbf{a}_{t+\tau}$. We assume that the synchronization offset τ may lie in an interval $[-s, s]$. First, we remove the first and last s frames from the audio signal and compute the audio features in the remaining cropped sequence of length $T - 2s$. Then, we compute the visual features $\mathbf{v}_t$, $t \in [1, 2s + 1]$ in all groups of $T - 2s$ consecutive frames. Finally, CCA is applied between the set of cropped audio features $\mathbf{a}$ and each visual feature group $\mathbf{v}_t$. We select the optimal temporal gap such that the correlation between audio and visual features is maximized according to:

$$\tau = \arg\max_t \lambda_t - (s + 1), \tag{12}$$

Algorithm 2 Audio-visual synchronization

```
Input: Audio and video streams, time interval $[-s, s]$.
Output: Synchronization offset $\tau$.
Delete the first and last $s$ frames from the auditory signal.
Compute the audio features in the remaining $T - 2s$ instances of the audio stream.
for all groups of $T - 2s$ consecutive frames do
    Compute the visual features $\mathbf{v}_t, t \in [1, 2s + 1]$.
    Estimate the CCA between the cropped audio and the visual features $\mathbf{v}_t$
end for
Estimate the temporal offset $\tau$ according to Eq. (12).
```



(a) (b) (c)

Fig. 3: Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral.

where λ corresponds to the largest eigenvalue, which is associated with the maximization of the canonical correlation between the audio feature vector and each group of visual features, as the audio feature vector is slid over the visual features. The steps of the audio-visual synchronization algorithm are summarized in Algorithm 2.

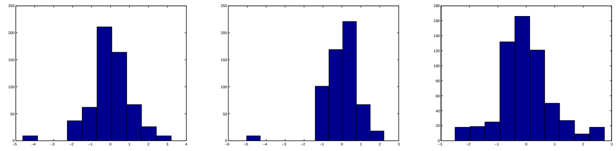
We now consider the fusion of the audio and visual features $\mathbf{a}$ and $\mathbf{v}$ respectively by projecting these features onto the canonical basis vectors $[\gamma_a^T, \gamma_v^T]^T$ and use this projection for recognition.

4 Experimental Results

In what follows, we refer to our *synchronized audio-visual* cues for *activity recognition* method by the acronym SAVAR. The experiments are applied to the novel *Parliament* dataset [19] and the TV human interaction (TVHI) dataset [21]. The number of features is kept relatively small in order not to increase the model's complexity.

4.1 Datasets

Parliament [19]: This dataset is a collection of 228 video sequences, depicting political speeches in the Greek parliament. All behaviors were recorded for 20 different subjects. The videos were acquired with a static camera and contain uncluttered backgrounds. The video sequences were manually labeled with one of three behavioral labels: *friendly* (90 videos), *aggressive* (73 videos), or *neutral* (65 videos). Figure 3 depicts some representative frames of the *Parliament* dataset. The subjects express their opinion on a specific law proposal and they adjust their body movements and voice intensity level according to whether they agree with that or not.



(a) (b) (c)

Fig. 4: Distribution of classes (a) friendly, (b) aggressive, and (c) neutral.

Each video sequence was manually labeled with one of three behavioral labels according to human perception on kindness and aggressiveness. The distribution of the three classes *friendly*, *aggressive*, and *neutral* is depicted in Figure 4. Each plot depicts the univariate histogram for each class. Note that all classes are not linearly separable.

The videos of the *Parliament* dataset were captured at a resolution of 320×240 pixels at 25 fps and their length is 250 frames. The dataset was annotated by two observers of Greek origin, who watched the videos independently and recorded their labels separately. Disagreement was resolved by a third observer. It is worth noting that the initial two annotators disagreed in only 3% of the videos of the dataset. The observers were asked to categorize the videos with respect to the notions of kindness and aggressiveness according to a general perception of a political speech by a citizen with a Greek mentality as follows. (i) Subjects with large and abrupt body, head and hand movements and high speech signal amplitude are to be labeled as aggressive. This corresponds to statesmen who express strongly their disagreement with the topic discussed or a previous speech given by a political opponent. (ii) Subjects with very small variations in their motion and speech signal amplitude are to be labeled as neutral. This class includes standard political speeches only expressing a point of view without any strong indication (body motion or voice tone) of agreement or disagreement with the topic discussed. (iii) Subjects with large but smooth variations in the pose of their body and hands speaking with a normal speech signal amplitudes are to be labeled as friendly.

TV human interaction [21]: This dataset consists of 300 video sequences collected from over 20 different TV shows. The video clips contain four kinds of interactions: *hand shakes*, *high fives*, *hugs* and *kisses*, which are equally distributed to the four classes (50 video sequences for each class). Negative examples (e.g., clips that do not contain any of the aforementioned interactions) consist the remaining 100 videos. The length of the video sequences ranges from 30 to 600 frames. The great degree of intra and inter-class diversity between the clips, such as different number of actors in each scene, variations in scale, and changes in camera angle, is an important factor that popularized this dataset for real world evaluation. Some representative frames of the TVHI dataset are illustrated in Fig. 5.

In particular, the *Parliament* and the TVHI datasets

Fig. 5: Sample frames from the TVHI dataset. (a) Hand shake, (b) High five, (c) Hug, and (d) Kiss.

are representative examples of individual and social behaviors, respectively. The *Parliament* contains examples of behavioral attributes, which may correspond to positive (e.g., friendliness) or negative (e.g., aggressiveness) behaviors. *Passive* is also a possible behavioral state for this dataset. The TVHI dataset on the other hand, models the social behaviors of people in terms of communication/interation with other people. Both kinds of behaviors entail much effort in order to analyze the given information.

4.2 Implementation details

We used 5-fold cross validation to split the *Parliament* dataset into training and test sets, and we report the average results over all the examined configurations. Moreover, for the same dataset, we also used the leave-one-speaker-out (LOSO) cross validation, to split training and testing data into two independent sets so that training and testing data may not have utterances from the same speaker. For the evaluation of our method to the TVHI dataset, we used the provided annotations, which are related to the locations of the persons in each video clip including the bounding boxes that contain them, the head orientations of each subject in the clips, the pair of the subjects who interact to each other and the corresponding labels. For comparison purposes, we used the same data split described in [21], which is a 10-fold cross validation. To obtain a bounding box of the human in every frame we used the method described by Dalal and Triggs [55]. Each frame is considered as a grid of overlapping blocks, where HOG features [50] are computed for each block. Finally, a binary SVM classifier is used to identify wether there exists an object or not. The detection window is extracted in all positions and scales and non-maximum suppression is used to detect each object. This method is able to cope with variations in appearance, pose, lighting and complex backgrounds.

The audio signal was sampled at 16 KHz and processed over 10 *ms* using a Hamming window with $25\,\%$ overlap. The audio feature vector consisted of a collection of 13 MFCC coefficients along with the first and second derivatives forming a 39 dimensional audio feature vector.

4.3 Model Selection

As shown in Fig. 2, there are many features that are non-informative due to pose variations or complex backgrounds. A comparison of the per class number of visual features before and after pruning using Algorithm 1 for both *Parliament* and TVHI datasets is illustrated in Fig. 6. It can be observed that the number of visual features before pruning is much higher than the number of visual features after pruning, which indicates that our pruning algorithm may significantly reduce the number of features by $29\,\%$ for the *Parliament* dataset and by $27\,\%$ for the TVHI dataset on average.

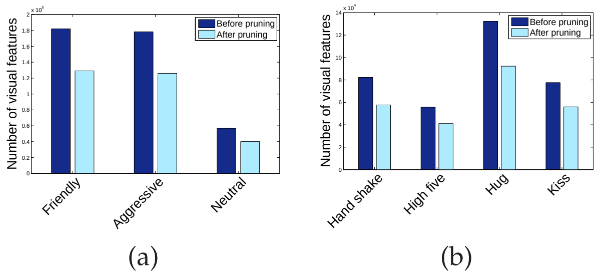


Fig. 6: Comparison of the per class number of visual features before and after pruning for (a) the *Parliament* and (b) the TVHI datasets.

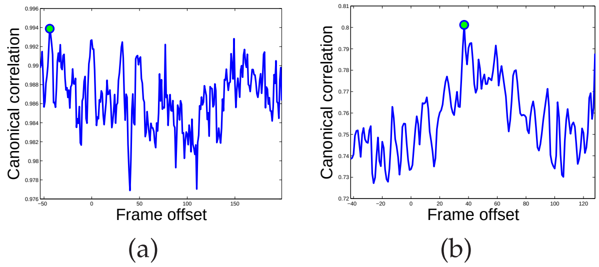


Fig. 7: Synchronization offsets between audio and video features for some sample video sequences of the *Parliament* (left) and TVHI (right) datasets. The circle indicates a delay of (a) -44 frames, (b) +37 frames.

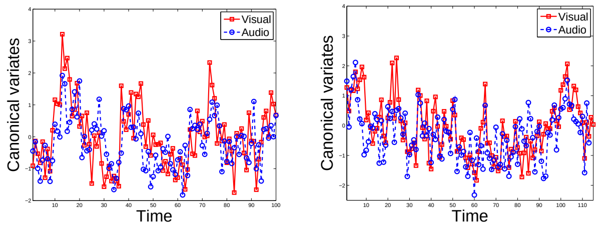


Fig. 8: Canonical variates of audio and visual features for two sample videos of the *Parliament* (left) and the TVHI (right) datasets. Notice the high correlation between audio and visual features obtained by the projection.

To automatically estimate the synchronization offset, such that the correlation between audio and video features is maximized, we used Algorithm 2. Figure 7 illustrates the synchronization offset for some randomly selected video sequences by plotting the most significant canonical basis as the visual features slide over the audio features. It is worth noting that, for the synchronization offset, we selected the frame with the maximum correlation. The corresponding canonical bases for the synchronized audio and visual features are depicted in Fig. 8. The similarity between the audio and visual canonical variates indicates high correlation.

The optimal number of hidden states was automatically estimated based on validation, varying the number of hidden states from three to ten. The L_2 regularization scale term σ was set to $10^k, k \in \{-3, \ldots, 3\}$. Finally, our model was trained with a maximum of 400 iterations for the termination of the LBFGS minimization method.

4.4 Results and Discussion

We compared the SAVAR approach, which uses audio-visual feature synchronization with an HCRF model, SAVAR(A/V sync), with previously reported methods in the literature and seven baseline approaches (variants of the proposed method). First, we compared the proposed SAVAR method with an HCRF variant, which does not employ audio-visual feature synchronization prior to the fusion process, SAVAR(A/V no-sync). To show the benefit of audio-visual fusion and synchronization, we compared our SAVAR(A/V sync) method against two HCRF variants, which use only audio, SAVAR(audio), and only visual, SAVAR(visual), features as input, respectively. Moreover, we compared our method with a late fusion technique without using audio-visual synchronization as it is not necessary in late fusion. Information from each modality was learned separately by the HCRF model and then the resulting classification scores were used as input to an SVM model to fuse the results. The parameters of SVM were chosen using cross validation.

A conditional random field model, using four different variants, was also used as a baseline method, to demonstrate the effectiveness of the HCRF model to learn the hidden dynamics between the video clips of different classes. First, synchronized and unsynchronized audio-visual features were used as input to two CRF models comprising two different variants A/V sync CRF and A/V no-sync CRF, respectively. Finally, we trained two CRFs, one with only audio features (audio CRF) and one with only visual (visual CRF) features.

4.4.1 Feature Pruning

The classification accuracy with respect to the number of hidden states before and after feature pruning for both the 5-fold and the LOSO cross validation schemes for the *Parliament* dataset is shown in Table 2. It is clear that the model obtained by the proposed algorithm, which uses pruned features, leads to better classification accuracy compared to the model, which uses the un-pruned features for both cross validation schemes. This is due to the fact that the un-pruned visual features may contain outliers and decrease the recognition accuracy, as the redundant visual features may lead to false estimation of the synchronization offset. Although audio features may improve the overall accuracy of the proposed method, in the case of un-pruned features they do not provide any significant performance as visual features may dominate over the audio features. For LOSO cross validation, and in contrast to the 5-fold scheme, visual features perform better than audio as there exist no utterances from the same speaker, and thus model overfitting, due to existence of redundant information, may be prevented. It is worth mentioning that the accuracy difference between visual and audio cues may be due to the difference in number of features for each modality. The optimal number of hidden states for the 5-fold and LOSO cross validation schemes, which use only audio and only visual data, in the case where feature pruning is used, is six. For the A/V no-sync method the optimal number of hidden states is 10. The number of hidden states remains the same for the LOSO scheme. The optimal number of hidden states for the proposed A/V sync method for the 5-fold scheme is seven, while for the LOSO scheme increases to nine.

Also, Table 2 shows the classification results with respect to the number of hidden states when late fusion is applied. It can be seen that the proposed method yields better results than late fusion for both 5-fold and LOSO cross validation schemes. For more than seven hidden states, the results of the proposed method are notably higher than those obtained by late fusion. Although late fusion may work better than the proposed method for a small number of hidden states (3, 5, and 6) for 5-fold cross validation, and 6 hidden states for LOSO cross validation, it is evident that for the majority of number of hidden states the proposed method performs better. Furthermore, even when late fusion outperforms the proposed approach, the improvement is marginal with respect to the improvement obtained by the proposed early fusion approach versus the late fusion for the same number of hidden states. This can be inferred by the fact that the optimal number of hidden states for the proposed 5-fold cross validation scheme is seven and the recognition accuracy is almost 30% higher than corresponding the late fusion approach for the same number of hidden states. Also, for the LOSO cross validation scheme, the recognition accuracy of the proposed method is higher in seven out of eight cases. This might be due to the low number of dimensions that late fusion handles. The proposed method exploits context provided by all modalities and the gain obtained by early fusion corresponds to the synchronized audio-visual cues, as they may be complementary in time. Also, despite the fact that late fusion is a suitable approach for handling multi-modal data, where each modality can be learned separately and differently, we may loose inter-modality dependence, which is crucial for audio-video classification.

The dependence of the classification accuracy and the number of hidden states on the TVHI dataset for both pruned and un-pruned features is shown in Table 3. Note that the visual model, which uses the original un-pruned features, performs better than the proposed A/V sync method, which uses pruned visual features, for six and

TABLE 2: Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h={3 . . . 10}) for the *Parliament* dataset [19] using 5-fold and LOSO cross validation, before feature pruning and after feature pruning.

#Hidden states:	3	4	5	6	7	8	9	10
HCRF before feature pruning using 5-fold cross validation								
A/V sync	29.0	55.7	56.8	**64.5**	46.3	47.7	51.4	51.0
A/V no-sync	34.6	46.5	**55.4**	51.0	34.1	44.7	42.0	44.4
Visual	44.9	56.6	47.6	52.9	44.1	40.9	**60.8**	48.9
HCRF before feature pruning using LOSO cross validation								
A/V sync	67.8	**70.0**	42.1	52.8	51.8	34.4	35.5	66.5
A/V no-sync	37.1	43.7	47.1	33.4	50.1	44.7	40.9	**53.9**
Visual	**48.4**	31.4	47.6	36.4	43.0	43.2	42.6	43.6
HCRF after feature pruning using 5-fold cross validation								
A/V sync	88.1	95.2	85.7	80.2	**97.6**	95.2	90.5	92.9
A/V no-sync	63.9	66.9	64.4	71.0	69.8	73.8	72.3	**78.9**
Audio	58.2	71.0	**72.7**	72.7	54.7	67.1	69.6	67.3
Visual	**67.1**	57.2	48.2	67.1	15.1	44.9	44.0	59.9
HCRF after feature pruning using LOSO cross validation								
A/V sync	91.0	89.7	94.9	77.1	93.6	94.9	**97.4**	97.4
A/V no-sync	63.0	59.3	74.9	80.4	76.9	79.2	75.1	**89.7**
Audio	59.3	**63.0**	50.0	63.0	51.9	53.7	62.7	50.0
Visual	42.7	63.7	58.2	**65.6**	60.0	42.7	39.6	58.2
Classification accuracies using late fusion								
Late-fusion (5-fold)	**91.1**	84.4	89.6	82.9	69.6	72.6	71.9	68.9
Late-fusion (LOSO)	83.3	78.7	**83.9**	81.5	63.2	67.1	69.3	68.9

10 hidden states. This is because the additional visual features may act as outliers and affect the estimation of the true synchronization offset. We can observe that in the case of feature pruning the visual model requires seven hidden states to achieve the best classification accuracy. It can also be noted that the audio model achieves the best recognition result by using four hidden states. Although the recognition results for this model are affected by background noise, it is obvious that the combination with the visual information can significantly improve the recognition rate. The A/V no-sync method requires eight hidden states, while the proposed A/V sync method uses nine hidden states to reach the best recognition accuracy. The number of hidden states depends not only on the number of the classes in a specific dataset, but also on the variety of the features used.

Table 3 demonstrates also the classification results, when late fusion is applied. Although in three out of seven cases, the late fusion scheme was able to improve the classification results, the proposed early fusion method performed better for the majority of the different number of hidden states. This is due to the heterogeneity of the different modalities and the confidence scores of each classifier, which may affect the discriminative ability of the SVM classifier as it may assign larger weights to scores that are less prominent.

TABLE 3: Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h={4 . . . 10}) the TVHI dataset [21] before feature pruning and after feature pruning.

#Hidden states:	4	5	6	7	8	9	10
HCRF before feature pruning							
A/V sync	40.6	**60.9**	46.9	43.8	53.1	54.7	54.7
A/V no-sync	39.1	42.2	40.6	32.8	46.9	**51.6**	35.9
Visual	35.9	37.5	48.4	42.2	29.9	35.9	**60.9**
HCRF after feature pruning							
A/V sync	53.1	79.7	70.3	73.4	73.4	**81.3**	76.6
A/V no-sync	46.9	53.1	35.9	56.6	**60.9**	54.7	42.2
Audio	**35.9**	34.4	29.7	28.1	28.1	32.8	23.4
Visual	28.1	50.0	59.4	**60.9**	37.5	35.9	57.8
Classification accuracies using late fusion							
Late-fusion	**80.1**	75.0	73.4	75.0	71.8	78.1	76.5

Taking a closer look at the visual model, we can see that the number of hidden states plays a crucial role in the recognition process; when the hidden states are increased from six to seven, recognition accuracy falls drastically from 67.1 % to 15.1 % for the *Parliament* dataset and from 60.9 % to 37.5 % for the TVHI dataset. In order to estimate the optimal number of hidden states we used cross validation. The reason for reporting the classification accuracies for all hidden states and not only for the optimal configuration is to demonstrate the behavior of the method with respect to the different number of hidden states and the cross validation schemes. It is also worth noting that 5-fold and LOSO cross validation schemes do not achieve the best accuracy for the same number of hidden states, which leads us to the conclusion that knowing in advance the optimal number of hidden states is not an easy task. Moreover, for both datasets, the optimal number of hidden states for each method with respect to the recognition accuracy is depicted in bold in Tables 2 and 3. When the same accuracy is achieved for more than one hidden states, the smallest number is considered to be the optimal. However, a larger number of hidden states may lead to a severe overfitting of the model. In this case, the regularization term in Eq. (9) may act as a preventer however, tuning the regularization parameters may be difficult and thus, overfitting may not be perfectly eliminated. It is also worth mentioning that both the *Parliament* and the TVHI datasets hold strong intra-class variabilities as certain classes are often confused because the subject performs similar body movements. This confirms that audio and visual information combined together constitute an important cue for action recognition.

4.4.2 Comparison of Learning Frameworks

Tables 4 and 5 report the classification accuracy on the *Parliament* dataset, for both 5-fold and LOSO cross validation schemes, and the TVHI datasets, respectively. We compare our SAVAR(A/V sync) method with the seven baseline methods and include previous results for each dataset reported in the literature. The results indicate that our approach captures the hidden dynamics between the clips (i.e., the interaction between an arm lift and the raise in the voice). It is clear that HCRFs

TABLE 4: Classification results on the *Parliament* dataset [19].

Method	**Accuracy (%)**			
	Audio	**Visual**	**A/V no-sync**	**A/V sync**
Vrigkas *et al.* [19]	N/A	**85.5** ± 0.412	N/A	N/A
SVM [40]	53.2 ± 0.053	65.7 ± 0.140	69.8 ± 0.135	72.6 ± 0.043
CRF [20]	50.3 ± 1.416	78.1 ± 1.560	67.6 ± 0.491	83.7 ± 0.653
SAVAR-5-fold	**72.7** ± 0.721	67.1 ± 0.389	78.9 ± 0.042	**97.6** ± 0.165
SAVAR-LOSO	62.2 ± 0.338	65.5 ± 0.347	**89.7** ± 1.613	97.4 ± 0.079

TABLE 5: Classification results on the TVHI dataset [21].

Method	**Accuracy (%)**			
	Audio	**Visual**	**A/V no-sync**	**A/V sync**
Patron-Perez *et al.* [21]	N/A	54.7	N/A	N/A
Li *et al.* [28]	N/A	**68.0**	N/A	N/A
Yu *et al.* [56]	N/A	66.2	N/A	N/A
Gaidon *et al.* [27]	N/A	55.6	N/A	N/A
Marín-Jiménez *et al.* [30]	**48.5**	46.0	54.5	N/A
SVM [40]	46.3 ± 0.008	56.7 ± 0.009	64.6 ± 0.012	75.9 ± 0.012
CRF [20]	36.7 ± 0.354	38.7 ± 0.527	49.5 ± 0.544	52.8 ± 0.746
SAVAR	35.9 ± 0.283	60.9 ± 0.028	**60.9** ± 0.644	**81.3** ± 0.191

outperform CRFs when multimodal data are used for the recognition task. Notably, our approach achieves very high recognition accuracy for the *Parliament* dataset (97.6 %), when 5-fold cross validation is used. Comparable results are also provided by the LOSO cross validation scheme as the recognition accuracy is only by 0.2 % lower than the 5-fold cross validation counterpart method. Note that for the SAVAR(A/V no-sync) variant, when LOSO scheme is used, the classification accuracy is by approximately 12 % higher than the corresponding 5-fold cross validation method. Also, when the 5-fold cross validation scheme is employed, SAVAR(audio) performs better than SAVAR(visual) as training data may have utterances from the same speaker. For the LOSO scheme, where the same speaker is excluded from the training data, visual features perform by approximately 3 % better than the acoustic.

The method in [19] employs a fully connected CRF model, where not only the labels but also the observation samples are associated to each other between consecutive frames. That is, the method in [19] assigns a distinct label to each frame, which makes it more suitable to cope with un-segmented videos (i.e., videos with more than one class labels). On the other hand, this property significantly increases the complexity of the method, which makes it quite difficult to use for large video clips.

Also, Table 5 demonstrates that the SAVAR approach performs significantly higher than other methods proposed in the literature for the TVHI dataset, by achieving an accuracy of 81.3 %, which is remarkably higher than the best recognition accuracy (68 %) for this dataset achieved by Li *et al.* [28], when only visual features are used, and the best recognition accuracy (54.5 %) achieved by Marín-Jiménez *et al.* [30], when audio and visual features are combined together. It is also worth noting that the SAVAR(visual) and the SAVAR(A/V no-sync) models achieve the same recognition accuracy for this dataset, indicating how important the audio-visual synchronization is for the recognition task, as the unsynchronized multimodal data may not provide any further information to the overall process. For the methods [21], [27], [28], [30], [56] the standard deviations of the classification accuracies are not provided in the original papers and thus, they are not included in Table 5.

TABLE 6: p-values of the proposed method for the *Parliament* dataset [19].

Method	**SAVAR-5-fold**	**SAVAR-LOSO**
Vrigkas *et al.* [19]	0.0200	0.0058
SVM [40]	0.0096	0.0001
CRF [20]	0.0137	0.0047

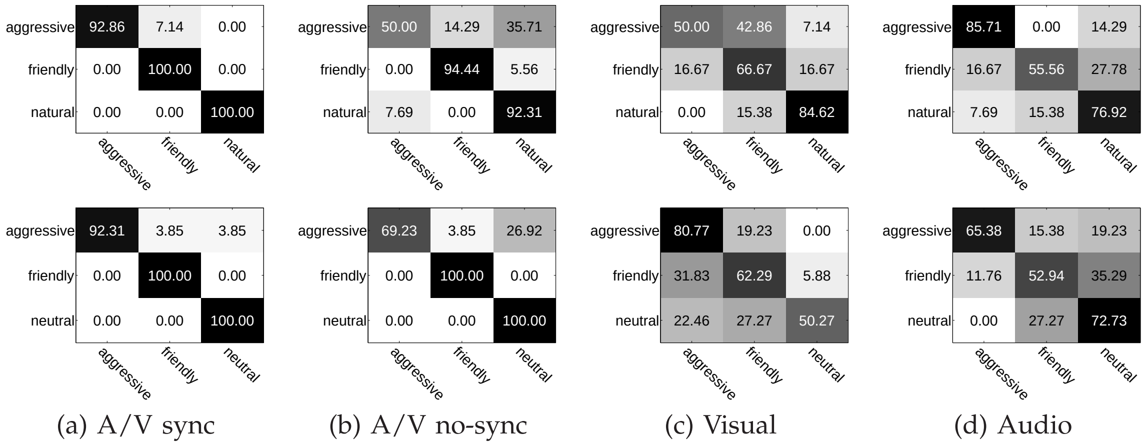
TABLE 7: p-values of the proposed method for the TVHI dataset [21].

Method	**SAVAR**
Patron-Perez *et al.* [21]	0.0012
Li *et al.* [28]	0.1239
Yu *et al.* [56]	0.0620
Gaidon *et al.* [27]	0.0015
Marín-Jiménez *et al.* [30]	0.0002
SVM [40]	0.0401
CRF [20]	0.0007

In order to provide a statistical evidence of the recognition accuracy, we computed the p-values of the obtained results with respect to the compared methods. The null hypothesis was defined as: the mean performances of the proposed model are the same as those of the state-of-the-art methods; and the alternative hypothesis was defined as: the mean performances of the proposed model are higher than those of the state-of-the-art methods. For the assessment of the statistical significance, we used paired t-tests with statistical significance threshold $p < 0.05$ for all experiments.

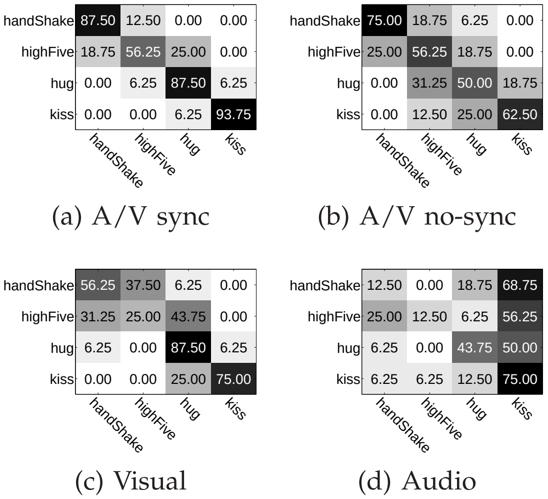
For the *Parliament* dataset (Table 6), we may observe that the SAVAR-5-fold and SAVAR-LOSO approaches reject the null hypothesis as all values are greater than the critical value (95% of significance level). For the TVHI dataset (Table 7) the null hypothesis is rejected for the majority of the cases. That is, for four out of seven cases the p-values were less than the significance level of 0.05. Therefore, we may conclude that the null hypothesis can be rejected and the improvements obtained by our model are statistically significant.

The resulting confusion matrices of the proposed method for the optimal number of hidden states for the *Parliament* dataset using 5-fold and LOSO cross validation, are depicted in Fig. 9. The proposed SAVAR(A/V sync) method has significantly small classification errors between different classes, when is compared to the other variants, for both 5-fold and LOSO cross validation schemes. The SAVAR(A/V no-sync) variant has also good classification results and particularly, for the LOSO cross validation scheme, it can perfectly recognize the



(a) A/V sync (b) A/V no-sync (c) Visual (d) Audio

Fig. 9: Confusion matrices for the classification results of the proposed SAVAR approach for the *Parliament* dataset [19], after feature pruning, using 5-fold cross validation (top row) and LOSO cross validation (bottom row).



(a) A/V sync (b) A/V no-sync

(c) Visual (d) Audio

Fig. 10: Confusion matrices for the classification results of the proposed SAVAR approach for the TVHI dataset [21], after feature pruning.

classes *friendly* and *neutral*. It is also interesting to observe that the different classes for the SAVAR(visual) and the SAVAR(audio) variants may be strongly confused, which emphasizes the fact that when combining audio and visual information together we are able to better separate the emotional states of a person.

Finally, the confusion matrices for the TVHI dataset are shown in Figure 10. The smallest classification error between classes belongs to the proposed SAVAR(A/V sync) method. Note that the different classes may be strongly confused as the TVHI dataset has large intra-class variability. Especially, the SAVAR(audio) variant has the largest classification error among all other variants as all classes are confused with the class *kiss*. This is due to the fact that in class *kiss* the audio information may serve as outlier since it contains background sounds.

The main strength of the proposed method is that it achieves remarkably good classification results when synchronized multimodal features are used compared with the results reported in the literature for the same datasets. Additionally, it keeps the number of visual features relatively small by pruning irrelevant features, thus reducing the computational burden of the method.

5 Conclusion

In this paper, we considered the problem of human behavior recognition in a supervised framework using a HCRF model with multimodal data. Specifically, we used audio features jointly with the visual information to take into account natural human actions. We proposed a feature selection technique for pruning redundant features, based on the spatio-temporal neighborhood of each feature in a video clip. This has helped reduce the number of features and sped up the learning process.

We also proposed a method for multimodal feature synchronization and fusion using CCA. We found that a moving subject is highly correlated with the auditory information, as human behaviors are characterized by complex actions of movements and sound emissions. The experimental results indicated that the exact synchronization of multimodal data before feature fusion ameliorates the recognition performance. In addition, the combination of audio and visual cues may lead to better understanding of human behaviors. The main strength of this method is that our multimodal fusion approach is general and it can be applied to several types of features for recognizing realistic human actions.

According to our results, the proposed SAVAR method, when it is used with synchronized audio-visual cues, achieves notably higher performance than all the compared classification schemes. This could be seen as an additional characteristic of our model to discriminate between similar classes, when multimodal data is used. Nonetheless, when only one modality was used, the method seemed to have difficulties in efficiently recognizing human behaviors, but it could yield comparable results to the multimodal SAVAR method. That is, although the combination of audio and visual cues could constitute a strong attribute for discriminating between different classes, each modality separately was unable to capture the variation in temporal patterns of the input data. The proposed method was also able to deal with natural video sequences. The visual feature pruning process could significantly reduce the amount of irrelevant features extracted in each frame, and considerably increased the classification performance with respect to all methods that do not incorporate feature pruning.

In the future, we plan to extend our model to cope with multimodal data, which can be considered mutually uncorrelated. Also, in the present work the number of hidden states is determined a priori. An automatic method necessitating more complex models is an issue of ongoing research.

Acknowledgments

This research was funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

[1] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, March 2010.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[4] Y. Song, L. P. Morency, and R. Davis, "Multimodal human behavior analysis: learning correlation and interaction across modalities," in *Proc. 14th ACM International Conference on Multimodal Interaction*, Santa Monica, CA, 2012, pp. 27–30.

[5] K. Bousmalis, L. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, March 2011, pp. 746–752.

[6] C. B. Germain and M. Bloom, *Human Behavior in the Social Environment: An Ecological View*. Columbia University Press, 1999.

[7] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, China, April 2013, pp. 1–8.

[8] N. Liu, E. Dellandréa, B. Tellez, and L. Chen, "Associating textual features with visual ones to improve affective image classification," in *Proc. Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Lisbon, Portugal, 2011, vol. 6974, pp. 195–204.

[9] M. S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," in *Proc. Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Lisbon, Portugal, 2011, vol. 6974, pp. 568–577.

[10] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *Proc. Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Lisbon, Portugal, 2011, vol. 6974, pp. 107–116.

[11] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, February 2013.

[12] S. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[13] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor., "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[14] Q. S. Sun, S. G. Zeng, Y. Liu, P. A. Heng, and D. S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, Dec. 2005.

[15] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for multimodal information fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2384–2387.

[16] T. Westerveld, A. P. de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processecing*, vol. 2003, pp. 186–198, Jan. 2003.

[17] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc 13th Annual ACM International Conference on Multimedia*, Singapore, 2005, pp. 399–402.

[18] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.

[19] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Classifying behavioral attributes using conditional random fields," in *Proc. 8th Hellenic Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 8445, May 2014, pp. 95–104.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 8th International Conference on Machine Learning*, San Francisco, CA, 2001, pp. 282–289.

[21] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, Dec. 2012.

[22] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 1226–1233.

[23] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, June 2013.

[24] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah, "Social cues in group formation and local interactions for collective activity analysis," in *Proc. 8th International Conference on Computer Vision Theory and Applications*, Barcelona, Spain, February 2013, pp. 539–548.

[25] X. P. Burgos-Artizzu, P. Dollár, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 1322–1329.

[26] L. P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007, pp. 1–8.

[27] A. Gaidon, Z. Harchaoui, and C. Schmid, "Recognizing activities with cluster-trees of tracklets," in *Proc. British Machine Vision Conference*, Surrey, UK, 2012, pp. 1–13.

[28] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier, "Activity recognition using dynamic subspace angles," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 3193–3200.

[29] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Los Alamitos, CA, 2009, pp. 58–65.

[30] M. J. Marín-Jiménez, R. M. noz Salinas, E. Yeguas-Bolivar, and N. P. de la Blanca, "Human interaction categorization by using audio-visual cues," *Machine Vision and Applications*, vol. 25, no. 1, pp. 71–84, 2014.

[31] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, "Semi-supervised multiple feature analysis for action recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 289–298, 2014.

[32] M. E. Sargin, Y. Yemez, E. Erzin, and A. M.Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[33] Q. Wu, Z. Wang, F. Deng, and D. D. Feng, "Realistic human action recognition with audio context," in *Proc. International Conference on Digital Image Computing: Techniques and Applications*, Sydney, Australia, December 2010, pp. 288–293.

[34] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008.

[35] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 4, pp. 875–885, 2013.

[36] Y. Song, L. P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, Jun. 2012.

[37] B. Siddiquie, S. M. Khan, A. Divakaran, and H. S. Sawhney, "Affect analysis in natural human interaction using joint hidden conditional random fields," in *Proc. IEEE International Conference on Multimedia and Expo*, San Jose, CA, July 2013, pp. 1–6.

[38] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011 -the first international audio visual emotion challenge," in *Proc. 1^{st} International Audio/Visual Emotion Challenge and Workshop*, ser. Lecture Notes in Computer Science, vol. 6975, Memphis, Tennessee, 2011, pp. 415–424.

[39] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[41] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, Lisbon, Portugal, 2007, vol. 4738, pp. 71–82.

[42] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, July 2014.

[43] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[44] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28^{th} International Conference on Machine Learning*, Bellevue, WI, 2011, pp. 689–696.

[45] A. Metallinou, M. Wollmer, A. Katsamani, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April 2012.

[46] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE International Conference onAcoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, pp. 3687–3691.

[47] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[48] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, May 2013.

[49] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2008.

[50] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. British Machine Vision Conference*, University of Leeds, Leeds, UK, September 2008, pp. 995–1004.

[51] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, September 2005.

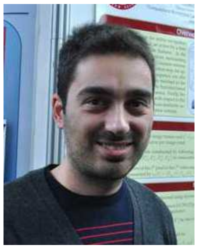
[52] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1996–2003.

[53] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412–424, March 2006.

[54] D. Mcennis, C. Mckay, I. Fujinaga, and P. Depalle, "jaudio: an feature extraction library," in *Proc. 6^{th} International Conference on Music Information Retrieval*, London, UK, September 2005, pp. 600–603.

[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, San Diego, CA, June 2005, pp. 886–893.

[56] G. Yu, J. Yuan, and Z. Liu, "Propagative Hough voting for human activity recognition," in *Proc. 12^{th} European Conference on Computer Vision*, 2012, pp. 693–706.



Michalis Vrigkas received the B.Sc. and M.Sc. in Computer Science from the University of Ioannina, Greece, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Engineering, University of Ioannina, Greece. Since January 2015 he is the Chair of the IEEE Student Branch of the University of Ioannina, Greece. His research interests include image and video processing, computer vision, machine learning and pattern recognition.

Christophoros Nikou received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1994 and the DEA and Ph.D. degrees in image processing and computer vision from Louis Pasteur University, Strasbourg, France, in 1995 and 1999, respectively. He was a Lecturer (2004-2009) and an Assistant Professor (2009-2013) with the Department of Computer Science and Engineering, University of Ioannina, Greece, where he has been an Associate Professor, since 2013. His research interests mainly include image processing and analysis, computer vision and pattern recognition and their application to medical imaging.

Ioannis A. Kakadiaris is a Hugh Roy and Lillie Cranz Cullen University Professor of Computer Science, Electrical & Computer Engineering, and Biomedical Engineering at the University of Houston. He joined UH in 1997 after a postdoctoral fellowship at the University of Pennsylvania. He earned his B.Sc. in physics at the University of Athens in Greece, his M.Sc. in computer science from Northeastern University and his Ph.D. at the University of Pennsylvania. He is the founder of the Computational Biomedicine Lab and the Director of the DHS Center of Excellence on Borders, Trade, and Immigration Research (CBTIR). His research interests include biometrics, video analytics, computer vision, pattern recognition, and biomedical image computing.