Expert Systems (2014). 31(4), 354-364.
DOI: http://dx.doi.org/10.1111/exsy.12040

Human action recognition with sparse classification and multiple-view learning

Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga and José M. Molina

Applied Artificial Intelligence Group, Universidad Carlos III de Madrid, Madrid, Spain
E-mail: rodri.cilla@gmail.com;rcilla@inf.uc3m.es

Abstract: *Employing multiple camera viewpoints in the recognition of human actions increases performance. This paper presents a feature fusion approach to efficiently combine 2D observations extracted from different camera viewpoints. Multiple-view dimensionality reduction is employed to learn a common parameterization of 2D action descriptors computed for each one of the available viewpoints. Canonical correlation analysis and their variants are employed to obtain such parameterizations. A sparse sequence classifier based on L1 regularization is proposed to avoid the problem of having to choose the proper number of dimensions of the common parameterization. The proposed system is employed in the classification of the Inria Xmas Motion Acquisition Sequences (IXMAS) data set with successful results.*

Keywords: Human Action Recognition, Multiple View Learning, L1 regularization

1. Introduction

The recognition of human actions has received increasing attention by the computer vision community during the last two decades (Lavee *et al.*, 2009; Weinland *et al.*, 2011). The objective is to build computational models to study how humans move and behave. A wide range of applications such as ambient intelligence (Chaaraoui *et al.*, 2012), video surveillance (Foresti *et al.*, 2004) or human–computer interaction (Ren & Xu, 2002) have benefitted from enhancements made in the field.

Whereas the first systems built for human action recognition were limited to a single camera (Cedras & Shah, 1995), advances made in visual sensor networks have motivated the inclusion of multiple cameras to enhance recognition robustness (Cilla *et al.*, 2012). This way, wider scenes can be covered, and systems can deal with occlusions caused by walls and furniture that complicate the recognition from a single camera view. Additionally, the perception of the motion from different viewpoints provides supplementary information that facilitates the recognition.

A current trend in human action recognition is the efficient combination of observations grabbed from different camera viewpoints. Most of the existing approaches are based on recovering a 3D model of the target of study exploiting multiple-view geometry. Visual hulls (Weinland *et al.*, 2006) or 3D skeletons (Parameswaran & Chellappa, 2006) are examples of these models. However, they have some drawbacks: (1) the need of accurate camera calibration parameters to recover the 3D model and (2) large network bandwidth requirements, as many raw data should be sent to a central node where the reconstruction is performed. All these facts make current 3D reconstruction methods not appropriate for their implementation in visual sensor networks.

Experimental evidence has shown that common 2D high-dimensional cues employed in human action recognition, such as silhouettes or optical flow, might be parameterized in low-dimensional manifolds where most of their variance is preserved (Blackburn & Ribeiro, 2007; Wang & Suter, 2008). At the same time, 3D descriptors recovered from multiple camera viewpoints, such as visual hulls or skeletons, are also parameterizable in low-dimensional manifolds (Turaga *et al.*, 2008; Peng *et al.*, 2009b). Both kinds of models contain information about the same real-world phenomena but at different depths, and both share the property of being parameterizable in low-dimensional manifolds preserving most of their variance. A question arising from this fact is if it would be possible to recover a low-dimensional manifold parameterization from multiple 2D descriptors with a similar performance to that recovered from 3D descriptors.

A design decision when learning low-dimensional manifold parameterizations is the number of dimensions that the learned low-dimensional space should have. There are different proposals to automatically adjust it. Variance preservance ratios (Jolliffe & MyiLibrary, 2002) or automatic relevance determination priors (Nounou *et al.*, 2002) are some popular choices. They adjust manifold dimensionality to preserve the maximum variance employing a reasonable number of parameters. However, these methods do not ensure a good performance when employing the obtained low-dimensional parameterization in a subsequent task, such as action class prediction. In practice, experimental cross-validation is employed to select the right number of dimensions. It has a high computational cost, as many set-ups have to be evaluated to find the best one. An alternative is to select the right number of dimensions in the subsequent task incorporating feature selection to the model with a sparse prior (Tibshirani, 1996).

This paper presents a multi-camera human action recognition system taking into account the considerations presented earlier. A feature fusion algorithm is proposed to learn a common manifold representation of feature

descriptors extracted from each camera viewpoint. The manifold is learned with a large number of dimensions. A sparse sequence classifier is proposed to select relevant features from the manifold parameterization. This way, the problem of choosing the right number of dimensions for the manifold vanishes.

1.1. Purpose and contributions

The purpose and contributions of this paper are summarized as follows:

- The aim of the paper is to provide a method to classify human actions using descriptors computed from different camera views.
- A feature fusion approach is proposed to combine action descriptors extracted from each camera view. Canonical correlation analysis (CCA) and kernel CCA (KCCA) are employed to obtain a common manifold parameterization of the motion descriptors.
- To avoid the problem of selecting the manifold dimensionality, a sparse hidden conditional random field (SHCRF) is introduced for action sequence classification. Sparsification is achieved by introducing an L1 penalty to the objective function optimized during training.
- An efficient online learning method is employed to train SHCRFs, reducing training time.
- The proposed system is evaluated in the recognition of IXMAS data set. Experimental evidence shows that the proposed method has a performance similar to state-of-the-art proposals employing 3D models.

1.2. Paper organization

This paper is organized as follows: Section 2 reviews relevant work on human action recognition with a special focus on methods employing multiple camera views. Section 3 introduces feature fusion algorithms to recover common manifold parameterizations for the action descriptors extracted from each one of the camera views. Section 4 introduces an SHCRF model employed to predict human actions while selecting relevant features from the manifold parameterization. Experimental results are presented and discussed in Section 5. Finally, Section 6 summarizes the contributions of this work.

2. Related work

2.1. Human action recognition

Multiple works have been developed to bridge the semantic gap from pixel intensity values in image sequences to descriptions of human actions performed in them. Each work defines different steps to solve the correspondence, but, in general, the process might be split in two steps: (1) feature extraction and representation and (2) action class prediction.

The former step deals with the extraction and efficient encoding of features to describe motions of interest. Multiple features might be extracted for motion modelling. Parametric models have been fitted to the targets, and temporal moments of the recovered parameter values have been employed for recognition (Ribeiro & Santos-Victor, 2005). Recovering model parameters is a difficult task, so these approaches have been deprecated in favour of visual features. Appearance descriptors, such as silhouettes or skeletons, describe how the target looks like (Bobick & Davis, 2001). Local motion descriptors, such as optical flow, describe the apparent motion of the pixels, providing a very strong cue for action recognition (Efros *et al.*, 2003). The main disadvantage of these methods is their lack of robustness towards partial occlusions of the target. Local feature descriptors have been proposed to overcome this limitation, encoding temporal and spatial variations in the neighbourhood of pixels with spatio-temporal saliency properties (Laptev, 2005).

The latter step deals with the transformation of features into semantic descriptions. Sequence models capture temporal correlations among feature values and employ them to select the appropriate action label. It is possible to apply exemplar-based models with different distances measures to select action labels (Efros *et al.*, 2003; Blackburn & Ribeiro, 2007; Wang & Suter, 2008), but most of the systems employ probabilistic graphical models. The generative Hidden Markov Model (HMM) is the de facto algorithm for human action recognition (Rabiner, 1989; Piccardi & Perez, 2007). However, discriminative graphical models have shown a better performance in action class prediction. The HCRF (Quattoni *et al.*, 2007) has outperformed HMMs in multiple action recognition tasks. However, there are many open problems related to the usage of HCRF. This work deals with model and feature selection for the HCRF.

2.2. Multi-camera approaches to human action recognition

Dasarathy's (1997) input–output data fusion model provides a categorization framework for data fusion systems according to the level of abstraction where system inputs and outputs are defined (*data*, features and decisions). Here, it is employed to organize relevant works to human action recognition from multiple camera viewpoints. Data-based levels of the framework are not considered, as the information employed in human action recognition is only defined at the feature and decision levels.

Diverse methods have been defined at the feature-in feature-out data fusion level to perform human action recognition from multiple camera viewpoints. It is possible to divide the works at this level into three different categories: (1) projection of 2D features to 3D; (2) feature fusion in a subspace; and (3) selection of the best view.

Different 3D representations might be obtained from projecting 2D features to 3D. A popular approach is to project 2D silhouettes to 3D to obtain a visual hull representing 3D appearance (Gkalelis *et al.*, 2009; Peng *et al.*, 2009a; Pehlivan & Duygulu, 2011). Visual hull reconstruction requires accurate silhouette segmentation from the different views. Recent works have proposed to project optical flow to 3D (Holte & Chakraborty, 2012) or to project local interest points (Holte *et al.*, 2011). Other works recover 3D star skeletons from 2D skeleton feature

correspondences (Chen *et al.*, 2008). Action sketch correspondence across multiple views has been also proposed (Yan *et al.*, 2008). The main drawback of 3D projection approaches is the need of accurate camera calibration parameters.

Other methods compute 2D features for each view and combine them by employing some simple scheme. Averaging of multiple features representing pose, global and local motion has been proposed, improving accuracy when compared with other alternatives (Maatta *et al.*, 2010). A joint bag-of-words histogram might be built with local feature descriptors extracted from each one of the camera views (Wu *et al.*, 2010), but a better performance is reported when other fusion strategies are employed. Projections maximizing cross-covariance have been learned to combine $\mathcal{R}$-transform derivatives extracted from each camera view (Karthikeyan *et al.*, 2011). Two-level linear discriminant analysis has been employed to learn silhouette projections maximizing action class separability (Iosifidis *et al.*, 2012). All these methods provide more flexible solutions for the combination of the features obtained from multiple cameras. However, experimental evidence shows a performance lower than that reported by 3D projection methods.

The last class of methods is based on the computation of a quality measure for each camera view, to perform the recognition employing only the data from the best view. An estimation of the orientation of human with respect to the camera (Shen *et al.*, 2007) or the different properties of the silhouette (Määttä & Aghajan, 2010) have been employed to obtain a quality measure. It has been proposed to select the camera with the highest number of detections (Wu *et al.*, 2010) when methods based on interest point location are employed. There are different quality measures for studying saliency, concavity or variations in silhouette stacks (Rudoy & Zelnik-Manor, 2011). The main drawback of these approaches is that they do not exploit complementary and redundant information obtained from multiple camera views.

Methods defined at the feature-in decision-out level encode existing correlations between feature descriptors extracted from each camera view and action labels. The concatenation of input features is the most straightforward procedure to perform data fusion in this way (Määttä & Aghajan, 2010; Wu *et al.*, 2010). The fused HMM (Wang *et al.*, 2007) proposes modelling correlations among observations coupling the values of the hidden state chains of parallel HMMs defined for each camera view. Histograms of local features are fused, rotating the ordering of the inputs to account for the variations in orientation (Srivastava *et al.*, 2009a). The main drawback of these works is their lack of flexibility, assuming that camera configuration remains unchanged between the train and test steps. A procedure to align camera views when the configuration changes from the train to test steps is defined in Ramagiri *et al.* (2011), but it requires to know relative camera placement.

Decision-in decision-out is the highest level of abstraction. Action prediction is performed for each camera view to later combine the results obtained. Majority voting is the most common method for the fusion of decisions (Määttä & Aghajan, 2010). A weighted voting strategy is proposed in Zhu *et al.* (2012), correcting each vote according to the value of the observed feature. Errors produced at each camera view might be incorporated into the voting procedure as proposed by Cilla *et al.* (2012).

3. Feature fusion for human action recognition

This section presents feature extraction and feature fusion methods employed in the proposed systems. The 2D human motion descriptor employed is first introduced. Then, CCA, a multiple-view dimensionality reduction method employed to find a common manifold parameterization of the motion descriptors, is described. This section finishes with a presentation of KCCA, a non-linear extension to CCA.

3.1. Feature extraction

Section 2 has shown that it is possible to employ different visual features in the recognition of human actions. This work employs the action descriptor proposed by Tran *et al.* (2008). It combines motion and appearance information. It has been selected because it has shown a high experimental performance in predicting the data set that will be employed for system evaluation.

It extracted normalizing the bounding box of the observed human to a square box preserving aspect ratio. Shape and optical flow are computed from the normalized box. Vertical and horizontal planes of the optical flow are split and blurred with a median filter. Thus, each box has three channels: silhouette, vertical flow and horizontal flow. The box is divided into four tiles, and a radial 18-bin histogram is computed from each tile and each channel. The obtained histograms are concatenated to obtain a 216-d vector. A Principal Component Analysis (PCA) reduction of the surrounding past, present and future vectors is appended to generate a descriptor of $d_{TRAN} = 286$ dimensions. Readers are referred to Tran *et al.* (2008) for more details.

3.2. Canonical correlation analysis

The objective of CCA (Hardoon *et al.*, 2004) is to find a pair of linear projections maximizing the correlation in the projected space between a pair of multivariate random variables. Given the zero-mean random variables in the input space x_1 and x_2 with dimensions d_1 and d_2, CCA finds a pair of linear transformations w_1, w_2 such that one component within each set of transformed variables is correlated with a single component in the other set. The correlation between the corresponding components is called canonical correlation, and there can be at most $d = \min(d_1, d_2)$ canonical correlations. The first canonical correlation is defined as

$$\rho = \max_{w_1, w_2} \frac{\langle w_1^T x_1 \cdot w_2^T x_2 \rangle}{\sqrt{\langle \| w_1^T x_1 \|^2 \rangle \langle \| w_2^T x_2 \|^2 \rangle}} \quad (1)$$

$$= \max_{w_1, w_2} \frac{w_1^T \langle x_1 x_2^T \rangle w_2}{\sqrt{w_1^T \langle x_1 x_1^T \rangle w_1 w_2^T \langle x_2 x_2^T \rangle w_2}} \quad (2)$$

where $\langle x_1 x_1^T \rangle$, $\langle x_2 x_2^T \rangle$ and $\langle x_1 x_2^T \rangle$ are estimated as $\widetilde{\Sigma}_{11}$, $\widetilde{\Sigma}_{22}$ and $\widetilde{\Sigma}_{12}$, respectively, that is, the different minors of the empirical covariance matrix $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix}$ of a set of training data $x = (x_1, x_2)$. The remaining canonical correlation directions are orthogonal to w_1 and w_2, respectively. They are solutions of the generalized eigenvalue problem:

$$\begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = (1+\rho) \begin{pmatrix} \widetilde{\Sigma}_{11} & 0 \\ 0 & \widetilde{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

The standard CCA model is defined for only two random variables x_1 and x_2. Bach and Jordan (2003) generalize CCA to m random variables. The generalized eigenvalue problem to solve is defined as follows:

$$\begin{pmatrix} \widetilde{\Sigma}_{11} & \cdots & \widetilde{\Sigma}_{1m} \\ \vdots & & \vdots \\ \widetilde{\Sigma}_{m1} & \cdots & \widetilde{\Sigma}_{mm} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} = \lambda \begin{pmatrix} \widetilde{\Sigma}_{11} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \widetilde{\Sigma}_{mm} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

where $\begin{pmatrix} \widetilde{\Sigma}_{11} & \cdots & \widetilde{\Sigma}_{1m} \\ \vdots & & \vdots \\ \widetilde{\Sigma}_{m1} & \cdots & \widetilde{\Sigma}_{mm} \end{pmatrix}$ denotes the empirical covariance matrix of a set of training data $x = (x_1, \ldots, x_m)$

3.3. Kernel canonical correlation analysis

The main limitation of the CCA model is given by the linearity of the projections obtained. Kernel methods (Burges, 1999) provide a procedure to transform linear algorithms based on inner products of the input data to non-linear algorithms, mapping the input data x to a high-dimensional feature space $\phi(x)$:

$$\begin{aligned} \phi : x &= (x_1, \ldots, x_n) \rightarrow \phi(x) \\ &= (\phi_1(x), \ldots, \phi_N(x)) \ (n < N) \end{aligned}$$

The linear algorithm (CCA in this case) is applied in the transformed feature space. The mapping from the input to feature spaces is not explicitly made. Instead, inner products performed by the linear algorithm in the input space are replaced by inner products in the feature space. Inner products in the feature space are computed by means of kernel functions. A kernel is a function K such that for all $x, z \in X$,

$$K(x, z) = < \phi(x) \cdot \phi(z) > \tag{3}$$

In CCA, data are introduced into the algorithm through the empirical covariance matrix $\widetilde{\Sigma}$. If the input data $X = [X_1 X_2]$ is centred, the minors of $\widetilde{\Sigma}$ are computed as follows:

$$\widetilde{\Sigma}_{11} = X_1^T X_1 \tag{4}$$

$$\widetilde{\Sigma}_{12} = X_1^T X_2 \tag{5}$$

Projection directions w_x, w_y can be replaced as the projection of the data into directions α_1 and α_2:

$$w_1 = X_1^T \alpha_1 \tag{6}$$

$$w_2 = X_2^T \alpha_2 \tag{7}$$

Equation (2) can then be rewritten as

$$\rho = \max_{\alpha_1, \alpha_2} \frac{\alpha_1^T X_1 X_1^T X_2 X_2^T \alpha_2}{\sqrt{\alpha_1^T X_1 X_1^T X_1 X^T \alpha_1 \cdot \alpha_2^T X_2 X_2^T X_2 X_2^T \alpha_2}} \tag{8}$$

Let $K_1 = X_1 X_1^T$ and $K_2 = X_2 X_2^T$ be the Gram matrices computed from the input data. Substituting into equation (8),

$$\rho = \max_{\alpha_1, \alpha_2} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{\alpha_1^T K_1^2 \alpha_1 \cdot \alpha_2^T K_2^2 \alpha_2}} \tag{9}$$

With this transform, the input data are now introduced into the algorithm through the Gram matrices K_1 and K_2 instead of the empirical covariance matrix $\widetilde{\Sigma}$. If the Gram matrices are obtained using a non-linear kernel function, the CCA algorithm now is non-linear. The generalized eigenproblem to solve in order to obtain the projection directions is

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

The eigenproblem in equation (10) has a trivial solution where all the values of ρ equal to one. That solution is not useful, and it can be avoided with a regularized version of the problem (Bach & Jordan, 2003):

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

$$= \rho \begin{pmatrix} \left(K_1 + \frac{N\kappa}{2} I\right)^2 & 0 \\ 0 & \left(K_2 + \frac{N\kappa}{2} I\right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

where N is the number of samples in the training set, κ is a small regularization constant and I is the $N \times N$ identity matrix.

Finally, the generalization of KCCA to m input variables is given by

$$\begin{pmatrix} \left(K_1 + \frac{N\kappa}{2}I\right)^2 & K_1K_2 & \dots & K_1K_m \\ K_2K_1 & \left(K_2 + \frac{N\kappa}{2}I\right)^2 & \dots & K_2K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_mK_1 & K_mK_2 & \dots & \left(K_m + \frac{N\kappa}{2}I\right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} =$$

$$\lambda \begin{pmatrix} \left(K_1 + \frac{N\kappa}{2}I\right)^2 & 0 & \dots & 0 \\ 0 & \left(K_2 + \frac{N\kappa}{2}I\right)^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \left(K_m + \frac{N\kappa}{2}I\right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

where K_m corresponds to the Gram matrix generated from the samples of the input variable m.

The kernel function employed in this paper is the Gaussian radial basis kernel:

$$K(x,\ z) \ = \ e^{-\frac{1}{\sigma^2}\|x\ -\ z\|} \tag{10}$$

where σ is a parameter controlling the bandwidth of the Gaussian.

4. Sparse sequence classification

This section presents the sequence classification algorithm employed to test the performance of the feature fusion method introduced in the previous section. A sparse version of the HCRF model is developed to select relevant features from the result of the feature fusion algorithm. The HCRF model is introduced in the first term, to later present the proposed sparse extension, and the optimization algorithm employed to recover optimal model parameters.

4.1. Hidden conditional random fields

The HCRF (Quattoni *et al.*, 2007) extends the CRF (Lafferty *et al.*, 2001) by introducing hidden state variables into the model. An HCRF is an undirected graphical model composed of three different sets of nodes, as shown in Figure 1. The node y represents the class label for an input sequence. $X = x_1,\ \dots,\ x_t$ is the set of nodes corresponding to the temporal observations in the input sequence. $H = h_1,\ \dots,\ h_t$ is the set of hidden variables modelling the relationship between the observations x_i and the class label y and the temporal evolution of the sequence.

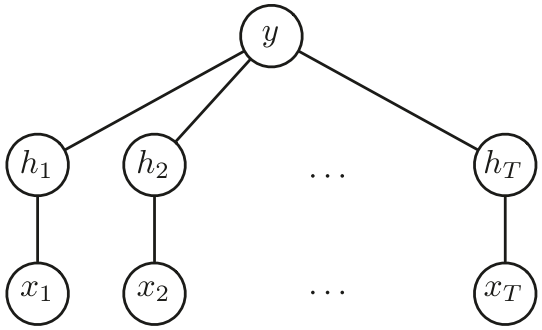


Figure 1: *Graphical model representation of the hidden conditional random field.*

The conditional probability of a sequence label y and a set of hidden part assignments $\mathbf{h}$ given a sequence of observations X is defined using the Hammersley–Clifford theorem of Markov random fields:

$$P(y, \mathbf{h}|\mathbf{x}, \theta) \ = \ \frac{e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y'}\sum_{h} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}} \tag{11}$$

where θ is the vector of model parameters. HCRFs belong to the general class of log-linear models. The conditional probability of the class label y given the observation sequence X is obtained by marginalizing over all the possible value assignments to hidden parts $\mathbf{h}$:

$$P(y|\mathbf{x},\ \theta)\ =\ \frac{\sum_h e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)}}{\sum_{y'}\sum_h e^{\Psi(y',\mathbf{h},\mathbf{x};\theta)}} \tag{12}$$

The potential $\Psi(y,\ \mathbf{h},\ \mathbf{x};\ \theta)$ is a linear function of the input variables:

$$\Psi(y,\mathbf{h},\mathbf{x};\theta)\ =\ \sum_i \phi(x_i)\ \cdot\ \theta(h_i)\ +\ \sum_i \theta(y,h_i) + \sum_{(j,\ k)\ \in\ E} \theta\left(y,h_j,h_k\right) \tag{13}$$

The first term, parameterized by $\theta(h_i)$, measures the compatibility of the observation at instant x_i with the assignment to the hidden variable h_i. The second term measures the compatibility of the hidden part h_i with the class label and is parameterized by $\theta(y,\ h_i)$. Finally, the third term models sequence dynamics, measuring the compatibility of adjacent hidden parts h_i and h_j with the class y.

Values for model parameters θ are estimated from training samples $\{x^i,\ y^i\}$ to maximize the L2-regularized conditional likelihood function of the model:

$$L(\theta)\ =\ \sum_{i\ =\ 1}^{n} \log \mathrm{P}(y_i|\mathbf{x}_i,\theta)\ +\ \frac{1}{2\sigma^2}||\theta||_2^2 \tag{14}$$

where the parameter σ controls the amount of penalization induced by the L2 norm of the model parameters. Different convex optimization techniques have been proposed to find the optimal parameters θ^* maximizing the conditional likelihood function in equation (14). Among them, Limited Memory - Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) (Liu & Nocedal, 1989) and conjugate gradient (Nocedal & Wright, 1999) are the most popular.

Inference of the posterior probability distribution in equation (12) and auxiliary probability distributions needed for the estimation of the conditional likelihood gradient in equation (14) is made, using belief propagation as proposed in Quattoni *et al*. (2007).

4.2. L1-regularized hidden conditional random fields

The L2-regularized training of the HCRF in equation (14) produces solutions $\theta*$ where all the components have a small value. This L2 regularization approach has some drawbacks from the computational learning perspective:

- There is no feature selection during training. Irrelevant features at input sequences are given a non-zero weight caused by the nature of the gradient of the L2 norm, producing model overfitting.
- No model selection is performed. Occam's razor principle of machine learning stands that the best model is the one with a lower complexity best adapting to training data. A consequence of not fulfilling this requirement is overfitting. Similarly to the previous case, parameters $\theta(y,\ h_i)$ and $\theta(y,\ h_j,\ h_k)$ corresponding to unnecessary hidden parts obtain a non-zero weight in the HCRF result. In practice, the right number of hidden parts is selected by means of cross-validation, requiring to test multiple configurations to find the best one.

A possible way to overcome these limitations is to replace the L2 penalty term in equation (14) by an L1 penalty term. This way, the objective function to minimize for parameter estimation has the form:

$$L(\theta)\ =\ \sum_{i\ =\ 1}^{n} \log \mathrm{P}(y_i|\mathbf{x}_i,\theta)\ +\ \frac{1}{\sigma}||\theta||_1 \tag{15}$$

The L1 norm causes components in θ to have a zero value if they are not needed. If a zero value is given to a parameter $\theta(h_i)$, then the corresponding input feature is not taken into account, producing a feature selection effect. If the zero value is given to a parameter $\theta(y,h_i)$ or $\theta(y,h_j,h_k)$, then model selection is performed, reducing the complexity of the model. Although the L1 regularization has been employed to estimate the optimal parameters of log-linear models such as CRFs (Lafferty *et al*., 2001), this is, to the best of our knowledge, the first time that it has been employed to train HCRFs.

Unfortunately, the L1 norm has the property of non-smoothness at 0, and conventional convex optimization techniques are no longer valid to recover optimal models parameters. Different approaches have been proposed for the optimization of L1-regularized log-linear models. A first approach is to reparameterize the model to transform the optimization problem into a smooth one (Vail *et al*., 2007). Model parameters are split in a pair of vectors $\theta\ =\ \theta^+\ -\ \theta^-$, such as $\theta^+\ >\ 0$ and $\theta^-\ >\ 0$. Standard convex optimization techniques are then applied to find the optimal parameters θ^*. Unfortunately, the convergence speed of the method is very slow requiring many iterations until an optimal solution is found.

An alternative to this method is to perform stochastic gradient descent (Tsuruoka *et al*., 2009). Stochastic gradient descent updates model parameters after presenting each sample. Although the obtained solutions are worse than those obtained using conventional methods, in practice, they are good enough to evaluate the performance of the feature fusion approaches proposed in the previous section.

The algorithm performs as follows:

$$\theta_i^{k\ +\ \frac{1}{2}}\ =\ \partial_i^k\ +\ \eta_k \frac{\partial L(j,\ \theta^k)}{\partial \theta_i^k} \tag{16}$$

$$\theta_i^{k\ +\ 1}\ =\ \begin{cases} max\left(0,\ \theta_i^{k\ +\ \frac{1}{2}}\ -\ \left(u_k\ +\ q_i^{k\ -\ 1}\right)\right) & \text{if } \theta_i^{k\ +\ \frac{1}{2}}\ <\ 0 \\ min\left(0,\ \theta_i^{k\ +\ \frac{1}{2}}\ +\ \left(u_k\ -\ q_i^{k\ -\ 1}\right)\right) & \text{if } \theta_i^{k\ +\ \frac{1}{2}}\ >\ 0 \end{cases} \tag{17}$$

where q_i^k is the total L1 penalty that θ_i has received up to the point:

$$q_i^k = \sum_{t=1}^{k} \left(\theta_i^{t+1} - \theta_i^{t+\frac{1}{2}} \right) \tag{18}$$

η_k controls the learning rate and at each iteration is decreased by an exponential decay:

$$\eta_k = \eta_0 \alpha^{\frac{k}{N}} \tag{19}$$

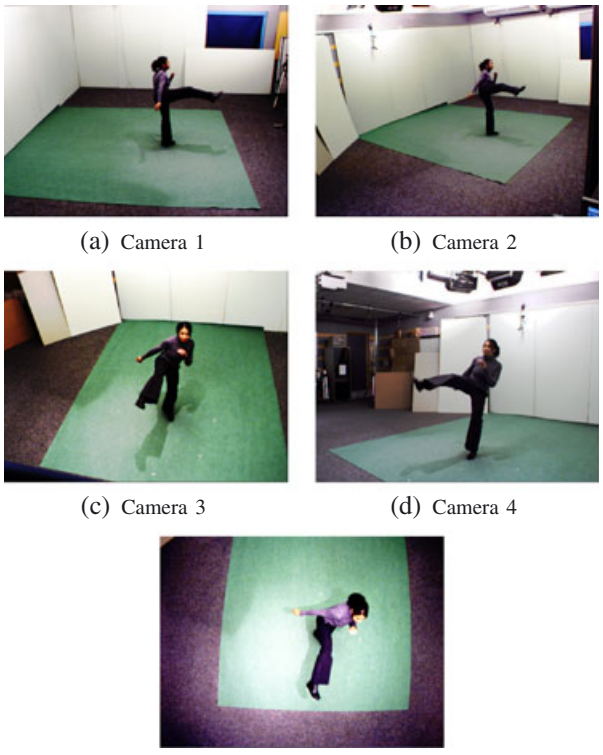
where η_0 and α are constants and N is the number of training samples.

5. Experiments

5.1. Experimental set-up

Algorithms presented in previous sections are going to be evaluated in the prediction of IXMAS data set (Weinland *et al.*, 2006). It contains 11 actions performed by 10 different actors at least three times each. The actions are recorded from five different viewpoints. The actions are (1) check watch; (2) cross arms; (3) scratch head; (4) sit down; (5) get up; (6) turn around; (7) walk; (8) wave; (9) punch; (10) kick; and (11) pick up. A sample frame of the data set observed from the five viewpoints is presented in Figure 2.

The protocol to evaluate system performance is the leave-one-actor-out cross-validation. The data set is divided into 10 different sets according to the actor performing each sample. The system is trained using all sets except one used for testing. The procedure is repeated until every actor has been used for testing.



(a) Camera 1 (b) Camera 2

(c) Camera 3 (d) Camera 4

(e) Camera 5

Figure 2: *The kick* action in the IXMAS data set from the five available views.

The CCA and KCCA are configured to provide 150 projection directions corresponding to the 150 highest canonical correlations. It should be emphasized again that a high value is given to this parameter as the SHCRF will select the relevant projections during training. KCCA is employed with a radial basis kernel of width $\sigma = 0.5$.

The SHCRF has been set up with $|H| = 22$ hidden parts. A value $\sigma = 0.2$ is given to the sparsity penalty term. The training algorithm is run for 40 iterations with a batch size of 20 samples. The exponential decay of the learning rate η is configured with an initial value $\eta_0 = 0.01$ and a decay rate $\alpha = 0.95$.

A Monte Carlo scheme is employed to evaluate the action prediction performance as the objective function employed to train the SHCRF is non-convex, producing different solutions depending on the initial value θ^0. The accuracy obtained for each test set is averaged over 30 different runs of the SHCRF training algorithm to obtain a real estimate of the algorithm performance.

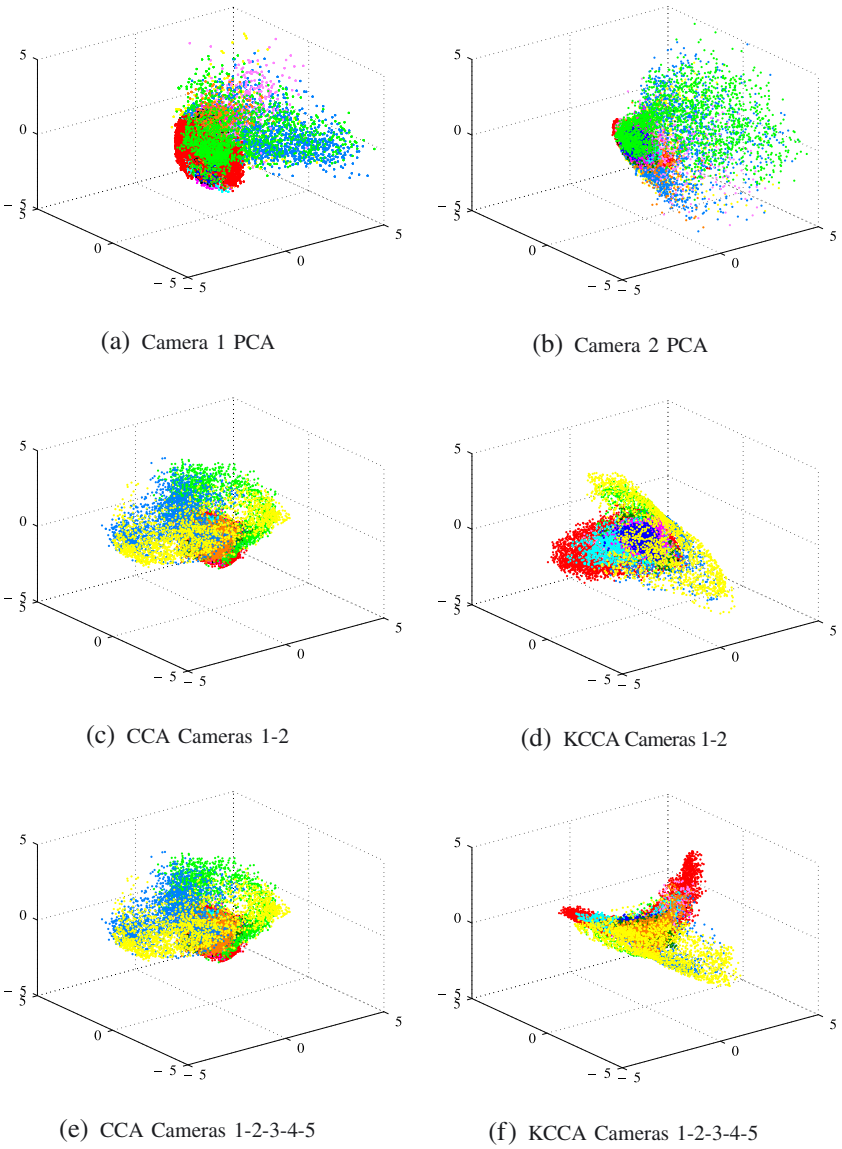
Finally, to measure the improvement produced by feature fusion in the predictive performance, a baseline model is going to be employed. The descriptors computed for cameras 1 and 2 are independently projected using PCA to retain the 150 most significant dimensions. The SHCRF is respectively evaluated using this projected sequences with the same procedure presented earlier.

5.2. Results and discussion

5.2.1. Feature fusion algorithms To visualize the effect of the feature fusion algorithms, the first three most significant features obtained by them and the PCA baseline are shown in Figures 3(a)–3(f). Projections have been obtained with the data of actors 2–10. It can be observed that fused features have a stronger class structure than features obtained by the PCA baselines, although they do not seem to be separated in any case. This is something normal in action recognition domains, as different action sequences usually share common frames, being their temporal evolution the real discriminative factor for action class prediction.

Another phenomenon observed in the plots of the linear models (Figures 3(a), 3(b), 3(c) and 3(e)) is that the main cause of class variation is produced by the direction and amount of movement. It can be seen that the variation of the actions sit down, get up and pick up, involving a vertical movement, seems to have a stronger structure than the others. However, in the features from the non-linear models presented in Figures 3(d) and 3(f), that behaviour is not present. Instead, class structure is appreciated for most of the actions. It seems that non-linear algorithms model other hidden factor of variations distinct from the movement direction.

5.2.2. Action classification Table 1 shows the accuracy obtained by the SHCRF predicting class labels for the IXMAS data set. For each evaluation fold, the worst, median and best results have been extracted. Results show that predicting action classes with fused features has an accuracy higher than predicting them with only one camera, although for CCA_{12} the worst case is worse than the classification with PCA_2.



(a) Camera 1 PCA (b) Camera 2 PCA

(c) CCA Cameras 1-2 (d) KCCA Cameras 1-2

(e) CCA Cameras 1-2-3-4-5 (f) KCCA Cameras 1-2-3-4-5

Figure 3: *Projections of the first three components obtained with the different dimensionality reduction algorithms. CCA, canonical correlation analysis; KCCA, kernel CCA.*

Table 1: *Worst, average and best case accuracy estimations of the different methods obtained using Monte Carlo leave-one-actor-out cross-validation*

	PCA_1	PCA_2	CCA_{12}	$KCCA_{12}$	CCA	KCCA
Minimum	0.634	0.688	0.685	0.709	0.748	0.748
Median	0.754	0.778	0.799	0.814	0.850	0.850
Maximum	0.865	0.868	0.880	0.907	0.934	0.940

Other interesting fact is that CCA and KCCA have almost the same accuracy when employing the data from the five camera viewpoints, whereas in the two cameras scenario, KCCA performs better than CCA. The feature fusion algorithms employed here extract common information in the data from each camera, so this saturation phenomenon seems reasonable. A non-linear algorithm extracts common information better than the linear algorithm, but as the number of data sources increases, the gap is reduced until they have a similar performance.

To better understand the behaviour of the presented algorithms, Figure 4 presents the box plots of accuracies obtained evaluating each actor following a Monte Carlo scheme. It should be noted that the feature fusion accuracy does not increase in all the cases reported. The most striking is when evaluating actor 2 (Figure 4(b)). The disastrous performance of camera 1 makes the fusion perform worse in all the cases than when using only camera 2. The cause for this phenomenon is that the fusion algorithms assume that every data source has a similar quality, not weighting them according to some quality metric. The phenomenon is produced for other actors too (Figures 4(h) and 4(i)), although there is not a big difference in the results using a single camera. That might be produced because the sources have complementary information that are not well represented in the transformed space.

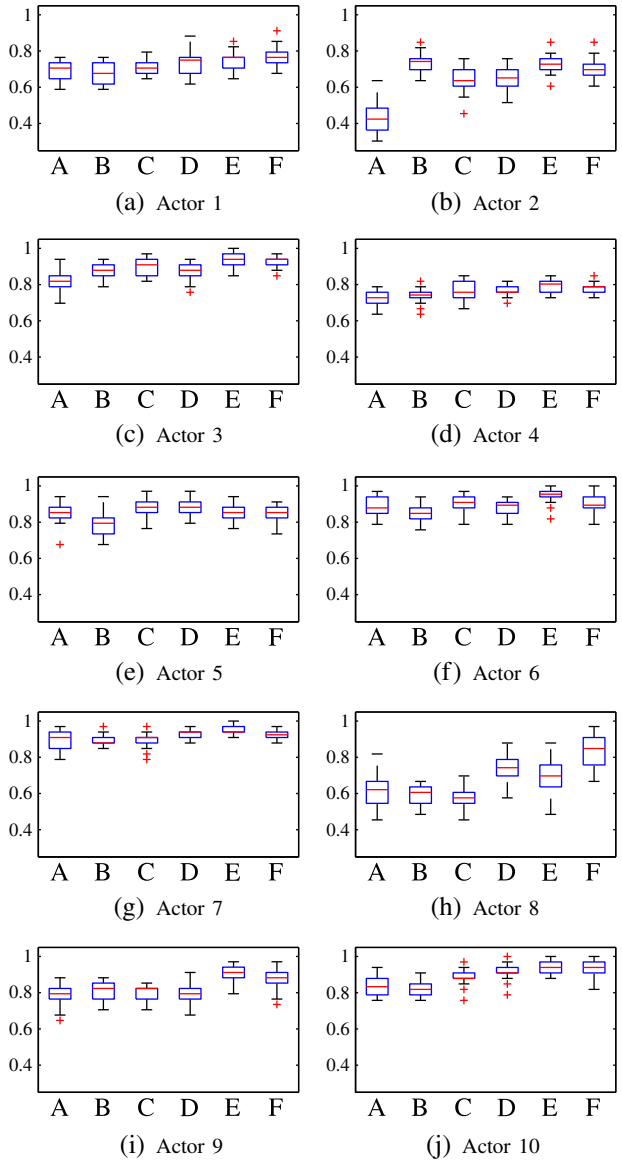


Figure 4: *Results for the different evaluation actors. A is the PCA from camera 1. B is the PCA from camera 2. C is the canonical correlation analysis (CCA) fusion of cameras 1 and 2. D is the kernel CCA (KCCA) fusion of cameras 1 and 2. E is the CCA fusion of the five cameras. F is the KCCA fusion of the five cameras.*

Finally, Table 2 compares the result of the presented proposal to others. The best case result performs as good as 3D proposals using visual hulls, whereas the average case improves results from decision-in decision-out.

Table 2: *Comparison of the accuracy of our method to others*

Method	Accuracy	Type
Srivastava *et al.* (2009b)	81.4	Decision-in Decision-out
Our Average	85.00	2D feature-in 2D feature-out
Weinland *et al.* (2006)	93.33	2D feature-in 3D feature-out
Our Best	94.00	2D feature-in 2D feature-out
Peng *et al.* (2009b)	94.59	2D feature-in 3D feature-out

6. Conclusions

This work has proposed the usage of multiple-view learning as a feature fusion method for human action recognition from multiple cameras. It has been shown that the usage of the information shared by rich 2D motion descriptors computed from multiple camera viewpoints improves the predictive accuracy. The usage of an L1-regularized sequence classifier has avoided the manual choice of the number of dimensions of the projected space. Testing multiple configurations to find the best dimension has been avoided, saving computational time. The usefulness of the proposed systems has been shown by predicting IXMAS data set with an accuracy similar to that reported by state-of-the-art methods.

References

Bach, F.R. and M.I. Jordan (2003) Kernel independent component analysis, *Journal of Machine Learning Research*, **3**, 1–48.

Blackburn, J. and E. Ribeiro (2007) Human motion recognition using isomap and dynamic time warping, in Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation, Rio de Janeiro, Brazil: Springer-Verlag, 285–298.

Bobick, A.F. and J.W. Davis (2001) The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 257–267.

Burges, C.J.C. (1999) Advances in Kernel Methods: Support Vector Learning, The MIT Press.

Cedras, C. and M. Shah (1995) Motion-based recognition: a survey, *Image and Vision Computing*, **13**, 129–155.

Chaaraoui, A.A., P. Climent-Pérez and F. Flórez-Revuelta (2012) A review on vision techniques applied to human behaviour analysis for ambient-assisted living, *Expert Systems with Applications*, **39**(12), 10873–10888.

Chen, D., P.C. Chou C.B. Fookes and, S. Sridharan (2008) Multi-view human pose estimation using modified five-point skeleton model. In International Conference on Signal Processing and Communication Systems 2007, 17–19 Dec 2007, Gold Coast, Australia.

Cilla, R., M.A. Patricio, A. Berlanga and J.M. Molina (2012) A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views, *Neurocomputing*, **75**(1), 78–87.

Dasarathy, B.V. (1997) Sensor fusion potential exploitation-innovative architectures and illustrative applications, *Proceedings of the IEEE*, **85**, 24–38.

Efros, A.A., A.C. Berg, G. Mori and J. Malik (2003) Recognizing action at a distance. *IEEE International Conference on Computer Vision*, **2**, 726–733.

Foresti, G.L., C. Micheloni and L. Snidaro (2004) Event classification for automatic visual-based surveillance of parking lots, in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. **3**, 314–317, IEEE.

Gkalelis, N., H. Kim, A. Hilton, N. Nikolaidis and I. Pitas (November 2009) The i3DPost multi-view and 3D human action/interaction database. *2009 Conference for Visual Media Production*, 159–168.

Hardoon, D.R., S. Szedmak and J. Shawe-Taylor (2004) Canonical correlation analysis: an overview with application to learning methods, *Neural Computation*, **16**, 2639–2664.

Holte, M.B. B. Chakraborty, J. Gonzalez and T. B. Moeslund (2012) A local 3D motion descriptor for multi-view human action recognition from 4D spatio-temporal interest points, *Selected Topics in Signal Processing, IEEE Journal of*, **6**(5), 553–565.

Holte, M.B., T.B. Moeslund, N. Nikolaidis and I. Pitas (May 2011) 3D human action recognition for multi-view camera systems. *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 342–349.

Iosifidis, A., A. Tefas, N. Nikolaidis and I. Pitas (2012) Multi-view human movement recognition based on fuzzy distances

and linear discriminant analysis, *Computer Vision and Image Understanding*, **116**, 347–360.
Jolliffe, I.T. (2002) Principal Component Analysis. Springer verlag.
Karthikeyan, S., U. Gaur, B.S. Manjunath and S. Grafton (November 2011) Probabilistic subspace-based learning of shape dynamics modes for multi-view action recognition. *2011 IEEE International Conference on Computer Vision Workshops* (*ICCV Workshops*), 1282–1286.
Lafferty, J., A. McCallum and F. Pereira (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data, in *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289.
Laptev, I. (2005) On space-time interest points, *International Journal of Computer Vision*, **64**, 107–123.
Lavee, G., E. Rivlin and M. Rudzsky (2009) Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video, *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, **39**, 489–504.
Liu, D.C. and J. Nocedal (1989) On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, **45**, 503–528.
Määttä, T. and H. Aghajan (2010) On efficient use of multi-view data for activity recognition, 158–165.
Määttä, T., A. Härmä and H. Aghajan (2010) On efficient use of multi-view data for activity recognition, in *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, (pp. 158–165), ACM.
Nocedal, J. and S.J. Wright (1999) Numerical Optimization, Springer Verlag.
Nounou, M.N., B.R. Bakshi, P.K. Goel and X. Shen (2002) Bayesian principal component analysis. *Journal of Chemometrics*, **16**, 576–595.
Parameswaran, V. and R. Chellappa (2006) View invariance for human action recognition, *International Journal of Computer Vision*, **66**, 83–101.
Pehlivan, S. and P. Duygulu (2011) A new pose-based representation for recognizing actions from multiple cameras, *Computer Vision and Image Understanding*, **115**, 140–151.
Peng, B., G. Qian and S. Rajko (August 2009a) View-invariant full-body gesture recognition via multilinear analysis of voxel data, *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras* (*ICDSC*), 1–8.
Peng, B., G. Qian and S. Rajko (September 2009b) View-invariant full-body gesture recognition via multilinear analysis of voxel data, *Third ACM/IEEE Conference on Distributed Smart Cameras*.
Piccardi, M. and O. Perez (2007) Hidden Markov models with kernel density estimation of emission probabilities and their use in activity recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR'07*), 1–8.
Quattoni, A., S. Wang, L.-P. Morency, M. Collins and T. Darrell (2007) Hidden conditional random fields, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**, 1848–1852.
Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**, 257–286.
Ramagiri, S., R. Kavi and V. Kulathumani (August 2011) Real-time multi-view human action recognition using a wireless camera network, *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 1–6.
Ren, H. and G. Xu (2002) Human action recognition in smart classroom, in *Automatic Face and Gesture Recognition, 2002. Proceedings of the Fifth IEEE International Conference on*, 417–422, IEEE.
Ribeiro, P.C. and J. Santos-Victor (2005) Human activity recognition from video: modeling, feature selection and classification architecture, in *International Workshop on Human Activity Recognition and Modeling* (*HAREM*).
Rudoy, D. and L. Zelnik-Manor (2011) Viewpoint selection for human actions, *International Journal of Computer Vision*, **97**, 243–254.
Shen, C., C. Zhang and S. Fels (2007) A multi-camera surveillance system that estimates quality-of-view measurement. *2007 IEEE International Conference on Image Processing*, III–193–III–196.
Srivastava, G., H. Iwaki, J. Park and A.C. Kak (August 2009a) Distributed and lightweight multi-camera human activity classification, *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras* (*ICDSC*), 1–8.
Srivastava, C., H. Iwaki, J. Park and A.C. Kak (September 2009b) Distributed and lightweight multi-camera human activity classification, in *Third ACM/IEEE Conference on Distributed Smart Cameras*, 1–8.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.
Tran, D., A. Sorokin and D. Forsyth (2008) Human activity recognition with metric learning, in Proceedings of the 10th European Conference on Computer Vision: Part I, Springer Berlin Heidelberg: Springer-Verlag, 561.
Tsuruoka, Y., J. Tsujii and S. Ananiadou (2009) Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: *Volume 1-Volume 1*, Association for Computational Linguistics, 477–485.
Turaga, P., A. Veeraraghavan and R. Chellappa (2008) Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision, in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8, IEEE.
Vail, D.L., J.D. Lafferty and M.M. Veloso (2007) Feature selection in conditional random fields for activity recognition, in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, 3379–3384, IEEE.
Wang, L. and D. Suter (2008) Visual learning and recognition of sequential data manifolds with applications to human movement analysis, *Computer Vision and Image Understanding*, **110**, 153–172.
Wang, Y., K. Huang and T. Tan (2007) Multi-view gymnastic activity recognition with fused HMM, In Computer Vision-ACCV 2007 (pp. 667–677). Berlin Heidelberg: Springer.
Weinland, D., R. Ronfard and E. Boyer (2006) Free viewpoint action recognition using motion history volumes, *Computer Vision and Image Understanding*, **104**, 249–257.
Weinland, D., R. Ronfard and E. Boyer (2011) A survey of vision-based methods for action representation, segmentation and recognition, *Computer Vision and Image Understanding*, **115**, 224–241.
Wu, C., A.H. Khalili and H. Aghajan (2010) Multiview activity recognition in smart homes with spatio-temporal features, *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras - ICDSC'10*, 142.
Yan, P., S.M. Khan and M. Shah (June 2008) Learning 4D action feature models for arbitrary view action recognition. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–7.
Zhu, F., L. Shao and M. Lin (2012). Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern Recognition Letters*, **34**(1), 20–24.

The authors

Rodrigo Cilla

Rodrigo Cilla received his BS, MS and PhD degrees in computer science from the Universidad Carlos III de Madrid in 2007, 2008 and 2012, respectively. His research interest includes, among others, human action recognition, multiple target tracking, high-dimensional data analysis, data fusion and biological image processing.

Miguel A. Patricio

Miguel A. Patricio received his BS and MS degrees in computer science and his PhD degree in artificial intelligence from the Universidad Politécnica de Madrid in 1991, 1995 and 2002, respectively. He has held an administrative position at the Computer Science Department of Universidad Politécnica de Madrid since

1993. He is currently an associate professor at the Escuela Politécnica Superior of the Universidad Carlos III de Madrid and a research fellow of the Applied Artificial Intelligence Group (GIAA). He has carried out a number of research projects and consulting activities in the areas of automatic visual inspection systems, texture recognition, neural networks and industrial applications.

Antonio Berlanga

Antonio Berlanga received the BA degree in physics from the Universidad Autónoma de Madrid, Spain, and the PhD degree in computer engineering from the Universidad Carlos III de Madrid in 1995 and 2000, respectively.
He is currently an associate professor with the Department of Computer Science, Universidad Carlos III de Madrid. His current research interests include evolutionary computation for multi-objective optimization, machine learning and data mining.

José M. Molina

José M. Molina is a full professor at the Universidad Carlos III de Madrid. He joined the Computer Science Department of the Universidad Carlos III de Madrid in 1993. Currently, he leads the Applied Artificial Intelligence Group (GIAA). His current research focuses on the application of soft computing techniques (neural network, evolutionary computation, fuzzy logic and multi-agent systems) to radar data processing, air traffic management and e-commerce. He is the author of more than 20 journal papers and 80 conference papers. He received his BS and PhD degrees in telecommunications engineering from the Universidad Politécnica de Madrid in 1993 and 1997, respectively.