This document is published in:

Manuel Ferrández, J. et al. (eds.), 2011. *Foundations on Natural and Artificial Computation: 4th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2011, La Palma, Canary Islands, Spain, May 30 - June 3, 2011. Proceedings, Part I*. Lecture Notes in Computer Science: 6686. Berlin, Germany: Springer-Verlag, pp. 491-500.

DOI: 10.1007/978-3-642-21344-1_51

Multicamera Action Recognition with Canonical Correlation Analysis and Discriminative Sequence Classification

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

Computer Science Department. Universidad Carlos III de Madrid
Avda. de la Universidad Carlos III, 22. 28270 Colmenarejo, Madrid, Spain
{rcilla,mpatrici}@inf.uc3m.es, {berlanga,molina}@ia.uc3m.es

Abstract. This paper presents a feature fusion approach to the recognition of human actions from multiple cameras that avoids the computation of the 3D visual hull. Action descriptors are extracted for each one of the camera views available and projected into a common subspace that maximizes the correlation between each one of the components of the projections. That common subspace is learned using Probabilistic Canonical Correlation Analysis. The action classification is made in that subspace using a discriminative classifier. Results of the proposed method are shown for the classification of the IXMAS dataset.

1 Introduction

The recognition of human actions has received an increasing attention by the computer vision community during the last years [10]. One of the current trends in the field is how to efficiently combine the perceptions grabbed from different viewpoints in order to create more robust action recognition systems. This way the system can cover wider scenes, being able to deal with the possible occlusions caused by walls and furniture that would make the recognition from a single view very difficult if not impossible.

Although there has been different proposals of human action recognition systems at the different sensor fusion levels proposed by Dasarathy [5] as [4,15] in the decision-in decision-out level or [22,12] at the feature-in decision-out level, the most successful approaches have been defined at the feature-in feature-out level. These approaches extract human silhouettes from the different cameras using for example background subtraction [16], and then reconstruct the 3D visual hull of the human [9] as the feature to be used for the recognition. This way, Weinland et al. [21] have proposed the Motion History Volumes (MHV) as an extension of the popular Motion History Image (MHI) [3] to 3D. Action classification is then made using Fourier analysis of the MHV. Peng et al [13] have performed multilinear analysis of the voxels in the visual hull. Turaga et al. [19] have studied the visual hulls using Stiefel and Grassman manifolds, reporting the best results for action recognition in 3D until date. The main drawback of these methods is that 3D visual hull reconstruction has a high computational

burden and requires accurate calibration parameters of each one of the cameras observing the scene. Also, the computation of the 3D visual hull requires at least the silhouettes from 2 different camera viewpoints.

This work presents a novel method for the recognition of human actions using multiple cameras at the feature fusion level but without explicitly reconstructing the visual hull or other 3D descriptor. Experimental results reported in the literature [19,13] have shown that visual hulls can be projected into low dimensional manifolds where most of their variance is preserved. Moreover, a silhouette is the projection of a visual hull into the camera plane, and different works [20] have also reported that they can be parametrized into low dimensional manifolds. The aim of our method is to find a set of projection functions, one for each camera, that project the corresponding silhouettes into a common low dimensional manifold. We think that the representation of the silhouettes into that common low dimensional manifold would be equivalent to the low dimensional representation fo the visual hull, so similar results can be achieved in human action recognition.

Probabilistic continuous latent variable models provide a framework for manifold learning where low and high dimensional representations are related via the factorization of their joint probability distribution. We test the usage of the Probabilistic Canonical Correlation Analysis (PCCA) model [2] to learn the projections of the features observed at the different cameras into a subspace that maximizes the correlation between their components. The representation of the observed features into that subspace is then used for action sequence classification.

Paper is organized as follows: section 2 presents how the proposed system is structured; 3 reviews the Canonical Correlation Analysis model; section 4 describes the sequence classifier that is going to be used to test the system; section 5 shows some experimental validation of the method; finally, section 4 discusses the conclusions and future lines of the work.

2 System Overview

The architecture of the proposed system is shown on figure 1. The images grabbed by the C different cameras observing the scene are independently processed to extract a sequence of action descriptors $X_c = x_1c, \ldots, x_{Tc}$, $1 \leq c \leq C$, and T is the total number of frames grabbed. The C sequences of actions descriptors extracted are fused projecting them into a common subspace to give a sequence of common action descriptors $Z = z_1, \ldots, z_T$, $z_t = F(x_{t1}, \ldots, x_{tC})$. Finally each sequence is introduced into an action classifier to make the decision on the action being performed in the sequence.

3 Canonical Correlation Analysis

Canonical Correlation Analysis is the method we use for the fusion of action descriptors. In the next paragraphs we give an overview of the classical and the probabilistic formulation.

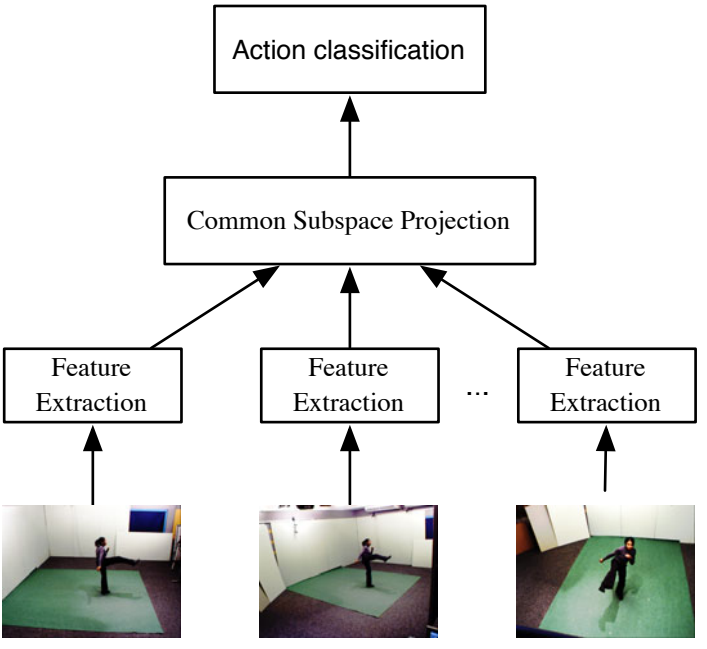


Fig. 1. Overview of the proposed system. Features are extracted for each available view. They are projected into a common subspace by Canonical Correlation Analysis. This projection is the used for action classification.

3.1 Classical Definition

Canonical Correlation Analysis [6] allows measuring the linear relationship between a pair of multidimensional variables. Given two random variables x_1 and x_2 of dimension d^1 and d^2 and zero mean, 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:

$$\begin{aligned}\rho &= \max_{w1,w2} \frac{\left\langle w_1^T x_1 \cdot w_2^T x_2 \right\rangle}{\sqrt{\left\langle \|w_1^T x_1\|^2 \right\rangle \left\langle \|w_2^T x_2\|^2 \right\rangle}} \\ &= \max_{w1,w2} \frac{w_1^T \left\langle x_1 x_2^T \right\rangle w_2}{\sqrt{w_1^T \left\langle x_1 x_1^T \right\rangle w_1 w_2^T \left\langle x_2 x_2^T \right\rangle w_2}}\end{aligned}$$

where $\left\langle x_1 x_1^T \right\rangle$, $\left\langle x_2 x_2^T \right\rangle$ and $\left\langle x_1 x_2^T \right\rangle$ are respectively estimated as $\tilde{\Sigma}_{11}$, $\tilde{\Sigma}_{22}$ and $\tilde{\Sigma}_{12}$, i.e, the different minors of the sample covariance matrix $\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{pmatrix}$

of a set of training data $x = (x_1, x_2)$. The rest of canonical correlation directions are orthogonal to w_1 and w_2 respectively. They can be computed as the solutions of the generalized eigenvalue problem:

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

The classical CCA model is defined for only two random variables x_1 and x_2. Bach and Jordan [1] generalize it to m random variables. The generalized eigenvalue problem to solve is then defined as:

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

3.2 Probabilistic Interpretation

Bach and Jordan [2] made a probabilistic interpretation of CCA extending the probabilistic interpretation of PCA proposed by Tipping and Bishop [17]. They define the following generative model, also shown on figure 2:

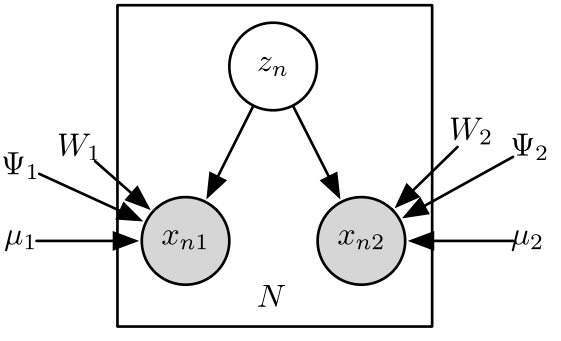


Fig. 2. Graphical model of the probabilistic interpretation of CCA made by Bach and Jordan [2] for two variables

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_q) \\ x_{n1} &\sim \mathcal{N}(W_1 z_n + \mu_1, \Psi_1) \\ x_{n2} &\sim \mathcal{N}(W_2 z_n + \mu_2, \Psi_2) \end{aligned}$$

They also show that the maximum likelihood estimates of the model parameters are given by:

$$\hat{W}_1 = \tilde{\Sigma}_{11} U_1 M_1 \tag{1}$$

$$\hat{W}_2 = \tilde{\Sigma}_{22} U_2 M_2 \tag{2}$$

$$\hat{\Psi}_1 = \Sigma_{11} - \hat{W}_1 \hat{W}_1^T \tag{3}$$

$$\hat{\Psi}_2 = \Sigma_{22} - \hat{W}_2 \hat{W}_2^T \tag{4}$$
$$\hat{\mu}_1 = \tilde{\mu}_1 \tag{5}$$
$$\hat{\mu}_2 = \tilde{\mu}_2 \tag{6}$$

where M_1 and M_2 are arbitrary matrices such that $M_1 M_2^T = P_q$, being P_q a matrix with the canonical correlations on its diagonal. U_{1q} and U_{2q} are the first q canonical directions.

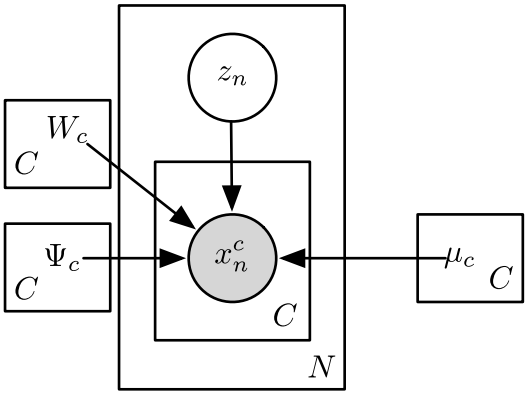


Fig. 3. Generalization of PCCA to C different data sources used in this paper

This model is easily generalizable to C different sources of data. The graphical model in that case corresponds to the shown on figure 3. Given a set of C different sources, each source c is generated as:

$$z_n \sim \mathcal{N}(0, I_q)$$
$$x_{nc} \sim \mathcal{N}(W_c z_n + \mu_c, \Psi_c)$$

The maximum likelihood estimates of the parameters are then given by:

$$\hat{W}_c = \tilde{\Sigma}_{cc} U_{cd} M_c \tag{7}$$
$$\hat{\Psi}_c = \Sigma_{cc} - \hat{W}_c \hat{W}_c^T \tag{8}$$
$$\hat{\mu}_c = \tilde{\mu}_c \tag{9}$$

This probabilistic generalization of the CCA model is employed in our system to combine the feature descriptors extracted from the different views. We choose the probabilistic interpretation as it would allow us to easily integrate the model as a part of larger graphical models for action recognition.

4 Hidden Conditional Random Fields

Hidden Conditional Random Fields (HCRF) [14] extend Conditional Random Fields [8] introducing hidden state variables into the model. A HCRF is an

undirected graphical model composed of three different set of nodes, as figure 4 shows. The node y represents the labelling of the input sequence. $X = x_1, \ldots, x_t$ is the set of nodes corresponding to the sequence observations $H = h_1, \ldots, h_t$ 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

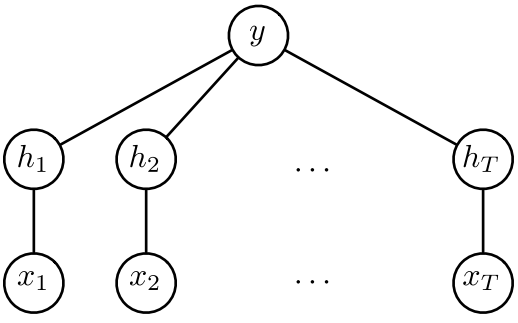


Fig. 4. Graphical model representation of the Hidden Conditional Random Field

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:

$$P\left(y, \mathbf{h} \mid \mathbf{x}, \theta\right) = \frac{e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y'} \sum_{h} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}} \tag{10}$$

where θ is the vector of model parameters. The conditional probability of the class label y given the observation sequence X is obtained marginalizing over all the possible assignments of hidden parts $\mathbf{h}$:

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

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

$$\begin{aligned} \Psi\left(y, \mathbf{h}, \mathbf{x}; \theta\right) = & \sum_{i} \phi\left(x_i\right) \cdot \theta(h_i) + \sum_{i} \theta\left(y, h_i\right) \\ & + \sum_{(j,k) \in E} \theta\left(y, h_j, h_k\right) \end{aligned} \tag{12}$$

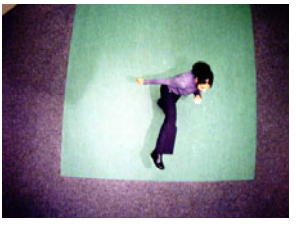
The first term, parametrized by $\theta\left(h_i\right)$ measures the compatibility of each observation x_i with the hidden variable h_i. The second term measures the compatibility of the hidden part h_i with the class label and is parametrized by $\theta((y, h_i)$. Finally, the third term models sequence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

(a) Camera 1

(b) Camera 2

(c) Camera 3

(d) Camera 4



(e) Camera 5

Fig. 5. The *kick* action in the IXMAS dataset from the five available views

Given a set of training samples $\{x^i, y^i\}$, model parameters are adjusted maximizing the L_2 regularized conditional likelihood function of the model:

$$L(\theta) = \sum_{i=1}^{n} \log P(y_i \mid \mathbf{x}_i, \theta) + \frac{||\theta||^2}{2\sigma} \tag{13}$$

The optimal parameters θ^* maximizing the conditional likelihood function are found using Quasi-Newton gradient based methods. Both the computation of the posterior probability on equation 11 and the auxiliary distributions that appear on the gradient of 13 can be efficiently made using belief propagation,as proposed in [14].

5 Experiments

5.1 Experimental Setup

The proposed algorithms are going to be tested in the classification of IXMAS dataset [21]. This dataset contains 11 actions performed by 10 different actors at least 3 times each. The actions are recorded from 5 different viewpoints. The algorithms are going to be tested using Leave-One-Actor-Out Cross Validation (LOAO-CV): The algorithms are trained with all the actors unless one, used for validation.

The system is going to be tested using the action descriptor proposed by Tran et al. [18], combining optical flow and appearance information. It is used in the system because it has shown a high experimental performance. The bounding box of a human being is normalized to a square box, from which human shape

and optical flow are computed. Vertical and horizontal planes of the optical flow are split and blurred. A radial histogram is computed over each of the optical flow planes and the shape. The three histograms are concatenated into 216-d vector. Lastly, PCA reduction of the surrounding past and future vectors is appended to finally generate a descriptor of $D_{TRAN} = 286$ dimensions. Readers are referred to [18] for more details.

In order to sped-up the CCA computation, PCA analysis of the descriptors is performed, retaining only the 100 principal components. CCA is going to be trained for latent dimensionality values of $q = 10, 12, 14, 16, 18, 205$. The number of hidden states of the HCRF is going to be fixed to $|H| = 11, 22$.

5.2 Results and Discussion

Table 1 shows the results obtained by the proposed method. It can be seen that the maximum accuracy is obtained for $q = 20$ dimensions and 22 hidden states. There accuracy starts growing with the number of dimensions, to then decrease and again increase to achieve teh best results. These phenomena gives an idea of how difficult is to manually parametrize the proposed methods

Table 1. Results obtained for different sizes of latent dimensionality

$\#$ dims / $\|H\|$	10	12	14	16	18	20
11	80.78	81.69	81.38	81.08	78.68	81.38
22	80.78	81.19	83.78	83.18	82.58	85.59

Finally, table 2 compares the results of our method to others. While it improves the results achieved by other methods, it is still far from the results obtained by the methods based on 3D visual hulls.

Table 2. Comparison of the accuracy of our method to others

Method	Accuracy	Type
Srivastava et al. [15]	81.4	Decision-in Decision-out
Our	85.59	Feature-in Feature-out
Weinland et al. [21]	93.33	2D Feature-in 3D Feature-out
Peng et al. [13]	94.59	2D Feature-in 3D Feature-out

6 Conclusions

This work has shown a preliminary approach to the fusion of features for human action recognition using a subspace learning technique. Feature descriptors extracted from different camera views have been projected into a common subspace learned using Canonical Correlation Analysis. The action classification has been

made in that subspace. Although the results achieved have been inferior to the obtained by state of the art 3D methods, we belief a non linear extension of the method using a mixture of Canonical Correlation Analyzers would reduce the gap [7].

Other strategy to test in the future would be to integrate the PCCA model into a sequence manifold learning method such the introduced in [11], in order to use the temporal evolution of the features for subspace regularization. Finally, other strategy to try would be to integrate the action classification with the resulting model, to perform the learning of the dimensionality reduction and the action classes at the same time.

References

1. Bach, F., Jordan, M.: Kernel independent component analysis. The Journal of Machine Learning Research 3, 1–48 (2003)
2. Bach, F., Jordan, M.: A probabilistic interpretation of canonical correlation analysis. Dept. Statist., Univ. California, Berkeley, CA, Tech. Rep 688 (2005)
3. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
4. Cilla, R., Patricio, M.A., Berlanga, A., Molina, J.M.: Fusion of single view soft k-NN classifiers for multicamera human action recognition. In: Corchado, E., Graña Romay, M., Manhaes Savio, A. (eds.) HAIS 2010. LNCS, vol. 6077, pp. 436–443. Springer, Heidelberg (2010)
5. Dasarathy, B.: Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proceedings of the IEEE 85(1), 24–38 (2002)
6. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
7. Klami, A., Kaski, S.: Local dependent components. In: Proceedings of the 24th international conference on Machine learning, pp. 425–432. ACM, New York (2007)
8. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (2001)
9. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 150–162 (1994)
10. Lavee, G., Rivlin, E., Rudzsky, M.: 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(5), 489–504 (2009)
11. Li, R., Tian, T., Sclaroff, S.: Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE, Los Alamitos (2007)
12. Määttä, T., Härmä, A., Aghajan, H.: On efficient use of multi-view data for activity recognition. In: 4th IEEE/ACM International Conference on Distributed Smart Cameras, ICSDC 2010, pp. 158–165 (2010)
13. Peng, B., Qian, G., Rajko, S.: View-Invariant Full-Body Gesture Recognition via Multilinear Analysis of Voxel Data. In: Third ACM/IEEE Conference on Distributed Smart Cameras (September 2009)

14. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(10), 1848–1852 (2007)
15. Srivastava, C., Iwaki, H., Park, J., Kak, A.C.: Distributed and Lightweight Multi-Camera Human Activity Classification. In: Third ACM/IEEE Conference on Distributed Smart Cameras, pp. 1–8 (September 2009)
16. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, Los Alamitos (2002)
17. Tipping, M., Bishop, C.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61(3), 611–622 (1999)
18. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 548–561. Springer, Heidelberg (2008)
19. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE, Los Alamitos (2008)
20. Wang, L., Suter, D.: Visual learning and recognition of sequential data manifolds with applications to human movement analysis. Computer Vision and Image Understanding 110(2), 153–172 (2008)
21. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104(2-3), 249–257 (2006)
22. Wu, C., Khalili, A., Aghajan, H.: Multiview Activity Recognition in Smart Homes with Spatio-Temporal Features. In: 4th IEEE/ACM International Conference on Distributed Smart Cameras, ICSDC 2010, pp. 142–149 (2010)