14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

HUMAN MODEL AND MOTION BASED 3D ACTION RECOGNITION IN MULTIPLE VIEW SCENARIOS

Cristian Canton-Ferrer, Josep R. Casas, Montse Pardàs

Image Processing Group,
Technical University of Catalonia
SUBMITTED TO MUSCLE SPECIAL SESSION

ABSTRACT

This paper presents a novel view-independent approach to the recognition of human gestures of several people in low resolution sequences from multiple calibrated cameras. In contraposition with other multi-ocular gesture recognition systems based on generating a classification on a fusion of features coming from different views, our system performs a data fusion (3D representation of the scene) and then a feature extraction and classification. Motion descriptors introduced by Bobick et al. for 2D data are extended to 3D and a set of features based on 3D invariant statistical moments are computed. A simple ellipsoid body model is fit to incoming 3D data to capture in which body part the gesture occurs thus increasing the recognition ratio of the overall system and generating a more informative classification output. Finally, a Bayesian classifier is employed to perform recognition over a small set of actions. Results are provided showing the effectiveness of the proposed algorithm in a SmartRoom scenario.

1. INTRODUCTION

Analysis of human motion and gesture in image sequences is a topic that has been studied extensively [1]. Detection and recognition of several human centered actions are the basis of these studies. The current paper addresses the problem of recognizing gestures of multiple persons in a SmartRoom in the framework of a motion and human model based analysis from multiple views. Multiple camera systems have been used for image and video analysis tasks in SmartRooms, surveillance, human-computer interfaces and scene understanding. From a mathematical viewpoint, multiple view geometry has been addressed in [9, 11], but there is still work to do for the efficient fusion of information from redundant camera views and its combination with image analysis techniques for object detection, tracking and higher semantic level analysis such as attitudes and behaviors of individuals.

Methods for motion-based recognition of human gestures proposed in the literature [1] have often been developed to deal with sequences from a single perspective [2, 4]. Considerably less work has been published on recognizing human gestures using multiple cameras. Mono-ocular human gesture recognition systems usually require motion to be parallel to the camera plane and are very sensitive to occlusions. On the other hand, multiple viewpoints allow exploiting spatial redundancy, overcome ambiguities caused by occlusion and provide 3D position information as well.

From an information processing perspective, most of the existing approaches to multiple view gesture recognition rely on information fusion at the feature level. This means that multiple inputs are separately analyzed to generate a motion description and then a classification of the gesture is performed on these data [2, 17]. This paper explores the complementary approach: first performing a fusion of the incoming data and then extracting 3D motion description features to perform classification.

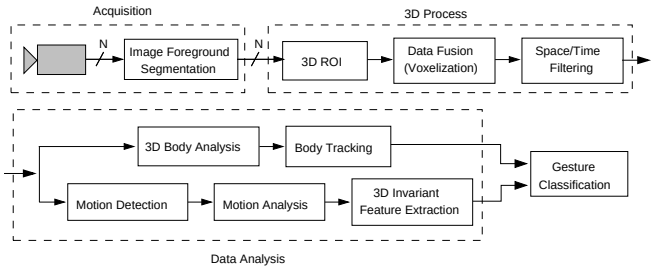


Figure 1: System flowchart: acquisition, 3D data processing, motion and body analysis and classification.

In [6], we introduced a method for 3D gesture recognition which is both robust to environmental conditions and computationally simple for real-time applications. Data fusion is achieved by exploiting redundancy among camera views to obtain a 3D representation of the scene. For the recognition of the movement, an extension of the motion representations proposed in [2] is presented: Motion History Volume and Motion Energy Volume. A set of robust 3D invariant statistical moments [15] are computed over motion information as a feature vector for classification. In this paper, we propose an extension of the previous action recognition technique by adding information regarding the position of the human body limbs. Taking into account that actions are produced by humans, an ellipsoid body model is fit to the incoming 3D data to capture in which body part the gesture occurs. This increases the recognition ratio of the overall system and generates a more informative classification output while still keeping the algorithm computationally simple for real-time purposes. Finally, motion and body model features are fed into a Bayesian classifier. Quantitative results for the proposed algorithm are provided as well as a comparison with other motion-based gesture recognition systems.

This method has been successfully applied to a multi-camera SmartRoom scenario in the framework of a scene understanding project involving recognition of human gestures in meetings. Other fields where our algorithm has potential applicability are interfaces for disabled people, body and gait analysis or domotics.

2. SYSTEM OVERVIEW

According to the flowchart depicted in Fig.1 the system comprises four data processing modules: image acquisition, 3D data processing, body and motion analysis and feature extraction, and classification.

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on ideal perspective projection. Accurate calibration information is available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's back-

ground learning and substraction technique [18, 13]. This adaptive algorithm proved to be robust when dealing with sequences with light changes in the background. It is assumed that the moving objects are human people. Segmented images, encoded as a binary mask, are the input information for the rest of image analysis modules described here since no color information is required.

2.1 3D Process Module

Prior to any further image analysis, the scene must be characterized in terms of space disposition and configuration of the foreground volumes, i.e. people candidates, in order to select those potential 3D regions where a gesture may appear. Images obtained from the multiple view camera system allow exploiting spatial redundancies in order to detect these 3D regions of interest. This task is carried out by the 3D processing module as explained below.

Once foreground regions are extracted from the set of N original images at time t, a set of M 3D points $\mathbf{x}^k$, $0 \leq k < M$, corresponding to the top most point of each 3D detected volume in the room is obtained by applying a robust Bayesian correspondence algorithm and tracking, as described in [5]. The information given by the established correspondences allows defining a Region of Interest (RoI) described by a bounding box $\mathcal{B}^k$, centered on each 3D top $\mathbf{x}^k$ with an average size adequate to contain a human candidate (see Fig.2(a)). This process allows reducing the complexity of the system discarding empty space regions not to be analyzed by forthcoming modules. For the sake of clarity, results presented in this article will refer to a single person in the scene while still being valid for multiple people.

As mentioned before, our approach to motion-based gesture recognition relies on feature extraction and classification over a fusion of the incoming information from the N cameras. Let us define a general fusion method from the data obtained by all N cameras at time instant t as the set

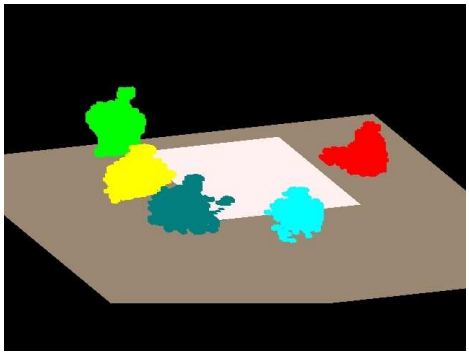
$$\Omega(\mathbf{x},t) = \left\{ I_n(\tilde{\mathbf{x}},t), \mathcal{B}^k(\mathbf{x},t), \mathcal{R}(\cdot) \right\} \qquad 0 \leq n < N, \quad (1)$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ state for 3D and 2D coordinates respectively, $I_n(\tilde{\mathbf{x}},t)$ is the segmented image captured by n-th camera, $\mathcal{B}^k(\mathbf{x},t)$ are the estimated volume RoIs and function $\mathcal{R}(\cdot)$ denotes the chosen data fusion procedure. In the current scenario where information present in the N images is originated by a common real 3D scene captured from different viewpoints, it is a sound assumption that a good data fusion process might be the reconstruction of the 3D scene itself. Other approaches to the 2D to 3D data fusion problem [3] generate new synthetic views by placing virtual cameras in an orthogonal coordinate system related with the center of the action. By working directly on the 3D result of the data fusion, our approach better captures the information available from the multiple views avoiding any redundancy on the data fed to the analyzer.

Taking the data provided by the foreground segmentation and the RoIs as input, reconstruction of 3D moving objects in the scene can be achieved by defining $\mathcal{R}(\cdot)$ as a robust Shape from Silhouette process [14]. This process generates a discrete occupancy representation of the 3D space (voxels) deciding whether a voxel is foreground or background by checking the spatial consistency of the N segmented silhouettes. Information derived from the multiple RoIs allows labeling the voxels as belonging to one person or another.

The data obtained with this 3D reconstruction is corrupted by spurious voxels introduced due to wrong segmentation, camera calibration inaccuracies, etc. The temporal analysis module placed next in the processing chain highly depends on the reliability of the data fusion. Isolated voxels should be removed not to be detected as motion. A connectivity filter is introduced in order to remove these voxels by checking its connectivity consistency with its neighbors in both space and time. An example of the output of the whole 3D processing module is depicted in Fig.2(b)

(a)



(b)

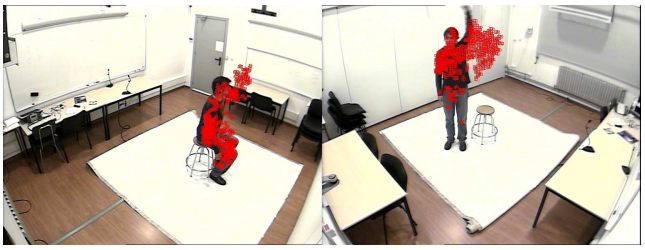
Figure 2: Example of the outputs from the 3D processing module in the SmartRoom scenario. In (a), multiview correspondences among regions of interest (RoIs) are correctly established. In (b), example of the data fusion set $\Omega(\mathbf{x},t)$ used in this paper.

2.2 Motion Analysis Module

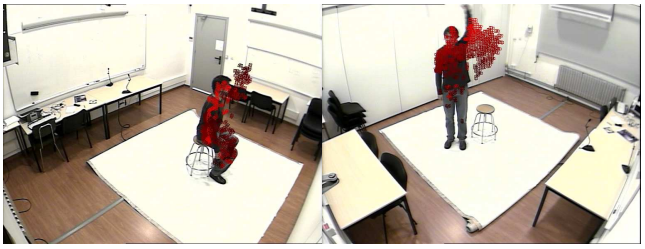
In order to achieve a simple and efficient low level view-dependent motion representation, [2] introduced the concept of Motion History Image (MHV) and Motion Energy Image (MEI). This representation has been recently used for monocular gait recognition tasks [10] and activity modeling [19]. We extend this formulation to represent view-independent 3D motion. In this way, ambiguities generated by occlusions are overcame. Analogously to [2, 4], the binary Motion Energy Volume (MEV) $E_\tau(\mathbf{x},t)$ is defined as:

$$E_\tau(\mathbf{x},t) = \bigcup_{i=0}^{\tau-1} \Omega^D(\mathbf{x},t-i), \quad (2)$$

where $\Omega^D(\mathbf{x},t)$ is the binary data set indicating regions of motion. This measure captures the 3D locations where there is motion in the last τ frames. Motion detection captured in $\Omega^D(\mathbf{x},t)$ can be coarsely estimated by a simple forward differentiation among voxel frames, still leading to satisfactory results while preserving a reduced computational complexity. It should be noted that τ is a crucial parameter in defining the temporal extent of a gesture. In Fig.3(a), an example of MEV is depicted.



(a)



(b)

Figure 3: Example of motion descriptors. In (a) and (b) are depicted the 2D projections of MEV and MHV respectively for gestures *sitting down* and *raising hand*.

To represent the temporal evolution of the motion, we define the Motion History Volume (MHV) where each voxel intensity is a function of the temporal history of the motion at that 3D location. Formally,

$$H_\tau(\mathbf{x},t) = \begin{cases} \tau & \text{if } \Omega^D(\mathbf{x},t) = 1 \\ \max\left[0, H_\tau(\mathbf{x},t-1) - 1\right] & \text{otherwise} \end{cases}. \quad (3)$$

This particular choice of temporal projection operator has the advantage that computation is recursive thus being a good representation for a real-time gesture recognition system. An example of MHV is shown in Fig.3(b).

Estimating a right value of the time factor τ (memory of the system) is critical to extract meaningful features to perform classification. Start and end of an action can be estimated adaptively by analyzing the volume activity of $\Omega^D(\mathbf{x},t)$: when there is an action starting, motion increases suddenly thus triggering the MHV computation until a gesture ends and motion activity decreases below a threshold A_{th} (see Fig.4).

2.3 Body Analysis Module

In order to extract a set of features describing the body of a person that performs an action, a geometrical configuration of human body must be considered. A number of body models have been proposed in the literature [1, 7] but most of them rely on computationally intensive minimization procedures to obtain valid body postures. Since the aim of our research is to increase robustness of gesture classification by embedding human body configuration information in our data analysis loop while keeping real-time performance, an ellipsoid model of human body has been adopted. In spite of this fairly simple approximation compared with more complex human body models, classification results proved the validity of our assumption as shown in Section 4.

Let $\mathcal{H} = \{\mathbf{c}, \mathbf{R}, \mathbf{s}\}$ be the set of parameters that define the ellipsoid modeling the human body candidate where $\mathbf{c}$ is the center, $\mathbf{R}$ the rotation along each axis centered on $\mathbf{c}$ and $\mathbf{s}$ the length of each axis. After obtaining the set of voxels $\Omega(\mathbf{x},t)$ describing a given person, we fit an ellipsoid shell to model it. Statistic moment analysis is employed to

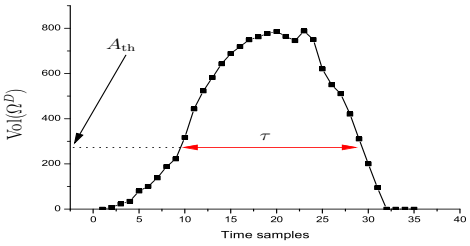


Figure 4: Estimation of time decay parameter τ of *hand waving* action by looking at the volume of the motion detection set $\Omega^D(\mathbf{x},t)$.

estimate the parameters of the ellipsoid from the centers of the marked voxels thus obtaining a 3D spatial mean $\bar{\Omega}$ and a covariance matrix $\mathbf{C}_\Omega$. The covariance can be diagonalized via an eigenvalue decomposition into $\mathbf{C}_\Omega = \mathbf{\Phi}\mathbf{\Delta}\mathbf{\Phi}^\top$, where $\mathbf{\Phi}$ is orthonormal and $\mathbf{\Delta}$ is diagonal. Identification of the defining parameters of the estimated ellipsoid $\mathcal{H}$ with moment analysis parameters is straightforward:

$$\mathbf{c} = \bar{\Omega}, \qquad \mathbf{R} = \mathbf{\Phi}, \qquad \mathbf{s} = \text{diag}(\mathbf{\Delta}). \quad (4)$$

This information is then fed to the body tracking module that refines this estimation taking into account body anthropometric restrictions imposing some motion and size constraints compatible with human bodies [7]. For example, the height of a person (largest value of matrix $\mathbf{s}$) restricts the possible locations of arms and legs according to the average lengths of body parts. Finally, time consistency of $\mathcal{H}$ parameters is achieved by a Kalman filter.

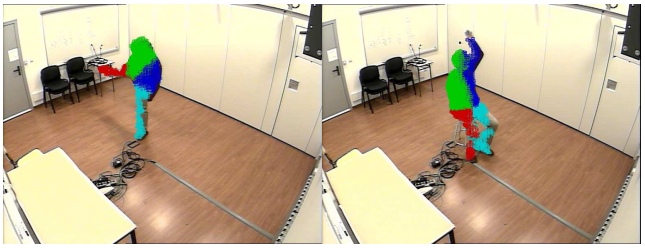
Once the parameters of the ellipsoid $\mathcal{H}$ representing the human body are computed, a simple body part classification can be derived. Voxels $\Omega(\mathbf{x},t)$ can be labeled as belonging to four categories: left/right-arm/leg (see Fig.5). These data will be used while performing classification of an action jointly with motion information.

3. FEATURE EXTRACTION AND GESTURE CLASSIFICATION

Data produced by the motion and body analysis modules is processed in order to extract a vector of features for classification.

Motion described at a low level using just image processing techniques requires a very high dimensional space to represent it. Methods to represent motion in a low-dimensional space are therefore desirable. Hence, informative features derived from the analyzed data (MHV and MEV in our case) are required. Statistical moments invariant to scaling, translation, rotation and affine mappings were early introduced by [12] for character recognition tasks. Their invariance properties yield to robust and informative features suitable for classification tasks and have been used in other 2D motion-based human gesture approaches [2, 4, 17]. The proposed system extends the usage of invariant moments to be computed over our data sets as classification features. Nevertheless, since our system is based on a data fusion prior to the classification process, 3D invariant statistical moments are required. These type of features have been already used in brain tissue classification tasks [16] and can be derived analytically. The reader is referred to Lo and Don's method [15] for a detailed description of the construction of invariant statistical moments of arbitrary dimension. For each data set $E_\tau(\mathbf{x},t)$ and $H_\tau(\mathbf{x},t)$, 5 invariant moment-based features are computed. Let us denote the set of these features as ψ_{MEV} and ψ_{MHV}.

(a)



(b)

Figure 5: Body analysis module output. In (a), original images for actions *quick* and *rise hand*. In (b), voxels belonging to the body of the person are labeled as belonging to right/left-arm/leg categories.

Information from body parts provided by body analysis module can be used to generate additional features. Extracted motion features do not capture any information regarding in which part of the body the motion has been produced. Let us call ψ_{BODY} the four features describing the relative amount of motion voxels located in each body part.

Given the computed moment-based motion features and the body features obtained for each of the actions to classify ω_j, $0 \leq j < K$, we define a full 14-dimensional feature vector as $\Gamma = [\psi_{\text{MEV}} \ \psi_{\text{MHV}} \ \psi_{\text{BODY}}]$. Even though the dimensionality of Γ is small, empty-space related problems arise when estimating class distributions [8]. Such effects decrease the efficiency of classification but this problem can be tackled by finding a transformed representation of data in a compact reduced dimensional space through Principal Component Analysis (PCA) [8]. By analyzing the training data we noticed that 90% of the variance of the data was achieved by using a dimension reduction to $d = 7$. Let us refer to the data set obtained after PCA analysis as $\hat{\Gamma}$.

The classification method is based on a Bayesian classification criterion assuming that $p(\hat{\Gamma}|\omega_j)$ is normally distributed. Since the noise in our data is the result of the sum of contributions from a large number of independent sources, Central Limit Theorem grants consistency to the Gaussianity assumption of our data. Indeed, further empirical tests [8] corroborate this assumption. Given an observation represented by $\hat{\Gamma}$, its classification is expressed by the maximum likelihood principle:

$$\arg\max_{\omega_j} p\left(\omega_j|\hat{\Gamma}\right), \tag{5}$$

where the posterior probability of a certain class ω_j given an observation $\hat{\Gamma}$ is formally

$$p\left(\omega_j|\hat{\Gamma}\right) = \frac{p\left(\hat{\Gamma}|\omega_j\right)\ p\left(\omega_j\right)}{p\left(\hat{\Gamma}\right)}. \tag{6}$$

Since $p(\omega_j)$ and $p(\hat{\Gamma})$ factors are wide and uninformative,

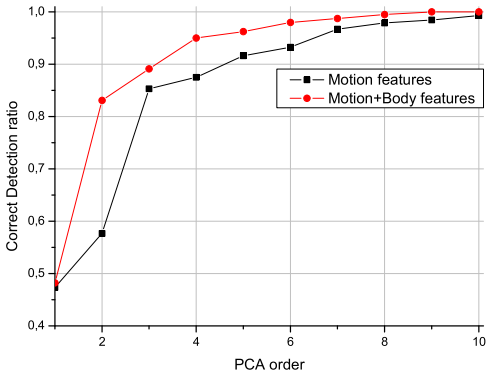


Figure 6: Classifier performance evaluated with motion and body features depending on the order of the PCA analysis.

Eq.6 can be expressed as

$$p\left(\omega_j|\hat{\Gamma}\right) \propto p\left(\hat{\Gamma}|\omega_j\right), \tag{7}$$

where $p(\hat{\Gamma}|\omega_j)$ is modeled as a multivariate Gaussian distribution defined by its mean μ and covariance matrix $\mathbf{\Sigma}$. Training data is used to estimate $(\mu, \mathbf{\Sigma})_j$ for each class in order to compute the class-likelihood discriminant in Eq.5.

4. RESULTS AND DISCUSSION

In order to evaluate the performance of the proposed algorithm, we collected a set of 70 training and 30 testing multi-view sequences of each action to be recognized. The analysis sequences were recorded with 5 fully calibrated and synchronized wide angle lenses cameras in the SmartRoom at UPC with a resolution of 768x576 pixels at 25 fps (see a sample in Fig.2(a)). The gesture category set was formed by 8 common actions of interest in the field of human-computer interfaces such as raising hand, sitting down, waving hands, crouching down, standing up, punching, kicking or jumping. Moreover, to show the effectiveness of our method and its robustness against rotations, occlusions and position, actions were recorded in different positions inside the room and facing various orientations.

Quantitative results shown in Table 1 prove the efficiency of the proposed algorithm to recognize human gestures from the given dataset. In average, we got a $p(\text{error}) = 0.0154$. Experiments have been carried out with and without these features to show the influence of body parts features in the overall performance. Fig.6 depicts the behavior of the classifier for diverse orders of the PCA analysis showing that body features increase the performance of the system.

Multiple view motion-based recognition of gesture is commonly addressed by the complementary information processing paradigm relying on feature fusion and classification. For general comparison purposes, we took the results provided in [17, 2] where the alternative approach to multiocular recognition of gestures is analyzed. Even though test databases are not the same, both contain similar actions. In comparison, the approach presented in this paper achieve lower error ratios. Moreover, our system has the advantage that no assumptions on the position and the orientation of the person are required due to the data fusion process. However, our method is conditioned by the initial foreground segmentation step thus being sensitive to the colors of the clothes of the people in the scene.

Table 1: Confusion matrix indicating the p(error) of the Bayesian classifier when using both motion and body features for classification.

	ω_0	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
ω_0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ω_1	0.0	-	0.0	**0.006**	0.0	0.0	0.0	0.0
ω_2	0.0	0.0	-	**0.010**	0.0	0.0	0.0	0.0
ω_3	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0
ω_4	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0
ω_5	0.0	0.0	0.0	0.0	**0.107**	-	0.0	0.0
ω_6	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0
ω_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-

5. CONCLUSIONS AND FUTURE WORK

We presented an efficient technique for motion-based view-independent gesture recognition in a multiple camera view environment. This paper explores the information processing methodology based on first performing a fusion of the incoming data and then extracting 3D motion description features. Classification is performed by jointly analyzing motion features and body position data obtained by fitting an ellipsoid body model.

Information provided by multiple views originated from the same real 3D world is better captured when being analyzed by a data-level fusion instead of a feature-level fusion. Experimental results proved the efficiency of our method proposing an alternative to the classical methodology to multi-ocular and mono-ocular motion-based gesture analysis [2, 17, 4]. Moreover, information regarding body parts position increases robustness of the overall system and generate a more informative classification output.

Future research within this topic involve developing more data fusion strategies involving color to generate informative descriptions of motion. More sophisticated classification techniques and 3D color related features are under research. The obtained data, allowing distinction among left/right and arm/leg as well as the classified action might be used by higher semantic analysis modules to analyze more complex and structured actions. Current research aims at using more detailed articulated body models that would lead to better classification results.

6. ACKNOWLEDGMENTS

This material is based upon work partially supported by the European Union under the NoE MUSCLE FP6-507752 and by the Spanish Ministry of Education under action ACERCA TEC2004-01914.

REFERENCES

[1] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: a review," in *Proc. IEEE Nonrigid and Articulated Motion Workshop*, 1997, pp. 90–102.

[2] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. on Pattern Anal. and Machine*, vol. 23, pp. 257–267, Mar. 1999.

[3] R. Bodor, B. Jackson, O. Masoud and N. Papanikolopoulos, "Image-Based Reconstruction for View-Independent Human Motion Recognition," in *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, Oct 2003, pp. 1548–1553.

[4] G. R. Bradski and J. W. Davis, "Motion Segmentation and Pose Recognition with Motion History Gradients," *Machine Vision and Applications*, vol. 13:3, pp. 174–184, Jul. 2002.

[5] C. Canton-Ferrer, J. R. Casas and M. Pardàs, "Towards a Bayesian Approach to Robust Finding Correspondences in Multiple View Geometry Environments," in *Proc. CGGM 2005*, LNCS, vol. 3515:2, pp. 281–289, May 2005.

[6] C. Canton-Ferrer, J. R. Casas and M. Pardàs, "3D Human Action Recognition in Multiple View Scenarios," submitted to *ICIP*, 2006.

[7] S. L. Dockstader, M. J. Berg and A. M. Tekalp, "Stochastic Kinematic Modeling and Feature Extraction for Gait Analysis," *IEEE Trans. on Image Processing*, vol. 12:8, pp. 962–976, Aug. 2003.

[8] R. Duda and P. Hart, *Pattern Classification*. John Wiley and Sons, 2001.

[9] O. Faugeras and Q. T. Luong, *The geometry of multiple views*. MIT Press, 2001.

[10] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Trans. on Pattern Anal. and Machine*, vol. 28:2, pp. 316–322, Feb. 2006.

[11] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[12] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Trans. on Information Theory*, vol. 8:2, pp. 179–187, Feb 1962.

[13] J. L. Landabaso, L. Q. Xu and M. Pardàs, "Robust Tracking and Object Classification Towards Automated Video Surveillance," in *Proc. Int. Conf. on Image Analysis and Recognition*, 2004, pp. 463–470.

[14] J. L. Landabaso, M. Pardàs and J. R. Casas, "Reconstruction of 3D Shapes Considering Inconsistent 2D Silhouettes," in *Proc. Int. Conf. on Image Processing*, to be published, 2006.

[15] C. Lo and H. Don, "3-D Moment Forms: Their Construction and Application to Object Identification and Positioning," *IEEE Trans. on Pattern Anal. and Machine*, vol. 11:10, pp. 1053–1063, Oct. 1989.

[16] J.F. Mangin et al, "Brain morphometry using 3D moments invariants," *Medical Image Analysis*, vol. 8:3, pp. 187–196, Aug. 2004.

[17] R. Rosales, "Recognition of Human Action Using Moment-Based Features," Tech. Report, Boston University, 1998.

[18] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Rec.*, 1999, pp. 252–259.

[19] T. Xiang and S. Gong, "Beyond Tracking: Modelling Activity and Understanding Behaviour," *Int. Journal of Computer Vision*, vol. 67:1, pp. 21–51, Feb. 2006.